RECONHECIMENTO DE AÇÕES COM HISTOGRAMAS DE CARACTERÍTICAS VISUAIS E CONTEXTO ADICIONADO POR TRANSFERÊNCIA DE APRENDIZAGEM

ANA PAULA BRANDÃO LOPES

RECONHECIMENTO DE AÇÕES COM HISTOGRAMAS DE CARACTERÍTICAS VISUAIS E CONTEXTO ADICIONADO POR TRANSFERÊNCIA DE APRENDIZAGEM

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo, Jussara Marques de Almeida (Co-Adv.)

Belo Horizonte

30 de setembro de 2011

ANA PAULA BRANDÃO LOPES

ACTION RECOGNITION WITH BAG-OF-VISUAL-FEATURES AND CONTEXTUAL INFORMATION ADDED BY TRANSFER LEARNING

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Arnaldo de Albuquerque Araújo, Jussara Marques de Almeida (Co-Adv.)

Belo Horizonte

September 30, 2011

Lopes, Ana Paula Brandão
L864r Action Recognition with Bag-of-Visual-Features and Contextual Information Added by Transfer Learning / Ana Paula Brandão Lopes. — Belo Horizonte, 2011
xxix, 123 f. : il. ; 29cm

Tese (doutorado) — Federal University of Minas Gerais
Orientador: Arnaldo de Albuquerque Araújo, Jussara Marques de Almeida (Co-Adv.)

1. Human Action Recognition. 2. Video Understanding. 3. Bag-of-Visual-Features. 4. Transfer Learning. 5. Context in Action Recognition. I. Título.

CDU 519.6*84 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Reconhecimento de ações com histogramas de características visuais e contexto adicionado por transferência de aprendizagem

ANA PAULA BRANDÃO LOPES

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ARNALDO DE ALBUQUERQUE ARAÚJO - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. JUSSARA MARQUES DE ALMEIDA - Co-orientadora
Departamento de Ciência da Computação - UFMG

PROF. EDUARDO ALVES DO VALLE JUNIOR
Faculdade de Engenharia Elétrica e Computação - UNICAMP

DR. IVAN LAPTEV
INRIA Paris - Rocquencourt

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. RICARDO DA SILVA TORRES
Instituto de Computação - UNICAMP

Belo Horizonte, 30 de setembro de 2011.

To Marcelo, Guilherme and Letícia. The love in my life, the salt of my life.

Acknowledgments

My dear friend Camillo Oliveira, having brought me to Belo Horizonte, is the first and most direct responsible to the mere existence of this work. For this, I thank Camillo more deeply than he can possibly imagine.

I also thank my advisor, professor Arnaldo. He provided me the support and also the freedom to explore my own ideas. In addition, he provided me with several opportunities to interact to as many colleagues as I could afford, from lab buddies to world-class researchers.

I thank professor Jussara Almeida, who kindly accepted to put her experienced and sharp mind available to me as a co-advisor. My experiments would for sure had been much more painful and less enlightening without her support.

I owe a lot to Eduardo Valle, great partner and mentor, advising me in many important ways. His generous sharing of experience and technical competence, together with his kind friendship were fundamental to have this work decently finished, despite all the hurdles.

I wholeheartedly thank Dr. Ivan Laptev, for his generosity in spending his time and sharing his expertise with us during his visit to Brazil. The discussions we had at that time provided a number of valuable insights, which influenced this thesis in a very definite way. I hope I have succeeded in making the best use of his efforts.

I also thank all the other teachers I was able to reach during my stay at PPGCC. There is at least one important lesson I have learnt from each one of you.

Sandra Avila was not only my first collaborator at NPDI, but also became a dear friend for me and my family. Júlia Epishina is another collaborator who became a dear and kind friend during this time. Together with Natália Bastista, these three great young women taught me a lot about hard work, strength and persistence.

Marcelo Coelho joined that core group a little, but quickly conquered everyone with his kindness, camaraderie and promptness to help in any matter, sometimes at his own expense.

The undergrads Rodrigo Oliveira and Elerson Santos are the true heros behind

a lot of the hard work which led to this thesis. Both were equally competent and trustworthy partners during all the time they worked with me. I owe Rodrigo a strong kick start to my work and Elerson a much needed smoothing of the final miles of this somewhat turbulent path.

I was blessed by being in a lab full of great, hard-working and kind people, and I have not interacted with all of them as much as I would like, but I am really grateful that I had known each one of these people.

The administrative staff of PPGCC is a rare group of helpful and generous professionals. I specially thank Renata, Sheila and Túlia, always available to provide all I needed to deal with a lot of paperwork with a minimum of obstacles.

Finally, I'm eternally grateful to my parents, Herodice and Olinda, who always supported and encouraged me, some times with more than they could really afford. To my beloved husband, Marcelo, who underwent a series of personal and professional sacrifices so I could follow my dream, which he fully embraced as *our* dream. It was him, together with my children, Guilherme e Letícia, who thankfully demanded me to keep at least one foot on the earth, even in the craziest of the deadlines eves. It is to them that I owe whatever slice of sanity that I was able to keep until the end of this journey.

"Mesmo quando tudo pede um pouco mais de calma
Até mesmo quando o corpo pede um pouco mais de alma
A vida não para.
(Even when everything asks for a little more calmness
Even when the body asks for a little more soul
Life does not stop). "
(Lenine/Dudu Falcão)

Resumo

Esta tese aborda a tarefa de reconhecimento de ações humanas em vídeos realísticos com base no conteúdo visual. Tal habilidade tem uma vasta gama de aplicações, mas este trabalho é motivado principalmente pela ideia de que descritores visuais e modelos efetivos são necessários para prover as máquinas de busca atuais com uma melhor capacidade de lidar com a grande quantidade de dados multimídia sendo produzidos todos os dias.

Uma questão levantada por estudos preliminares é o fato de que a coleta manual de exemplos de ações em vídeos é uma tarefa demorada e sujeita a erros. Isso é um gargalo importante para a pesquisa relacionada à compreensão de vídeos, uma vez que as grandes variações intra-classe apresentadas nestes vídeos exigem numerosos exemplos de treino.

Nesta tese, é proposta uma abordagem para esse problema baseada na teoria de Transferência de Aprendizagem, na qual é relaxada a suposição clássica de que os exemplos de treino e teste devem vir da mesma distribuição. Os experimentos com a base auxiliar Caltech256 e a base alvo Hollywood2 indicaram que com informações transferidas de apenas quatro conceitos da base auxiliar já é possível obter melhorias estatisticamente significativas na classificação das ações da Hollywood2. Esses resultados foram considerados uma evidência segura em favor da solução aqui proposta. Tal solução engloba nossa principal tese, que pode ser sumarizada em: a) é factível usar técnicas de Transferência de Aprendizagem para detectar conceitos em bases de vídeos realísticos de ações humanas, e b) usando a informação transferida, é possível melhorar o reconhecimento de ações em tais cenários.

Palavras-chave: Reconhecimento de Ações Humanas, Compreensão de Vídeos, Histogramas de Características Visuais Locais, Transferência de Aprendizagem, Contexto em Reconhecimento de Ações.

Abstract

This thesis addresses the task of recognizing human actions in realistic videos based on their visual content. Such an ability has a wide variety of applications in specific settings, but this work is above all motivated by the idea that effective visual descriptors and models need to be provided in order to make current search engines better able to cope with the large amount of multimedia data being produced every day.

An issue which has arisen from preliminary studies is the fact that to manually collect action samples from realistic videos is a time-consuming and error-prone task. This is a serious bottleneck to research related to video understanding, since the large intra-class variations of such videos demand training sets large enough to properly encompass those variations.

In this thesis, we propose an approach for this problem based on Transfer Learning (TL) theory, in which we relax the classical supposition that training and testing data must come from the same distribution. Our experiments with Caltech256 and Hollywood2 databases indicated that by using transferred information from only four concepts taken from the auxiliary database we were able to obtain statistically significant improvements in classification of most actions in Hollywood2 database, thus providing strong evidence in favor of the presented solution. Such solution encompasses our main thesis, which can be summarized in two main contributions: a) it is feasible to use TL techniques to detect concepts in realistic video action databases and, b) by using the transferred information, it is possible to enhance action recognition in those scenarios.

Palavras-chave: Human Action Recognition, Video Understanding, Bag-of-Visual-Features, Transfer Learning, Context in Action Recognition.

List of Figures

List of Tables

List of Acronyms

BoVF	Bag-of-Visual-Features	11
BoW	Bag-of-Words	15
CBVR	Content-Based Video Retrieval	36
DoG	Difference of Gaussians	12
DTM	Deep Transfer via Markov Logic	30
EMD	Earth's Mover Distance	55
ERS	Enhanced Random Selection	85
GMM	Gaussian Mixture Models	52
HCI	Human-Computer Interaction	57
HMM	Hidden-Markov Models	55
HoF	Histograms of optical Flow	14
HoG	Histograms of Gradients	14
IR	Information Retrieval	49
ITL	Inductive Transfer Learning	29
KTH	Royal Institute of Technology – in Swedish	50
LDA	Latent Dirichlet Allocation	51
LIBSVM	LIBrary for Support Vector Machines	25
LSCOM	Large-Scale Concept Ontology for Multimedia	55
MEI	Motion Energy Image	44
MHI	Motion History Image	44
ML	Machine Learning	28
MMI	Maximization of Mutual Information	52
MPEG	Moving Picture Experts Group	55

NMF	Non-negative Matrix Factorization	45
NN	Neural Networks	20
PCA	Principal Component Analysis	18
pLSA	probabilistic Latent Semantic Analysis	51
PRS	Pure Random Selection	85
RBF	Radial-Basis Function	25
ROC	Receiving Operator Characteristic	77
ROI	Region of Interest	37
RVM	Relevance Vector Machines	53
SIFT	Scale-Invariant Feature Transform	11
SFF	Smart Fast-Forward	47
SSM	Self-Similarity Matrix	56
ST-SIFT	Spatio-Temporal Scale-Invariant Analysis	
STIP	Space-Time Interest Points	13
STPM	Spatio-Temporal Pyramid Matching	52
SURF	Speed-Up Robust Features	13
SVM	Support Vector Machines	11
tf-idf	term-frequency inverse-document-frequency	19
TL	Transfer Learning	29
TTL	Transductive Transfer Learning	30
TREC	Text REtrieval Conference	
TRECVID	TREC Video Retrieval Evaluation	55
UTL	Unsupervised Transfer Learning	30
VWC	Video Words Clusters	52

Contents

Chapter 1

Introduction

This thesis addresses the problem of recognizing human actions of videos out of their visual content. An initial study on the different approaches for this problem found in the literature led to the choice of Bag-of-Visual-Features (BoVF) representations for the videos, with actions being modeled and classified by Support Vector Machines (SVM). From this basis, we implement and test an approach in which contextual information is used to improve the classical BoVF representation. Context is inferred from occurrences of a number of concepts throughout the video segment. Finally, those concepts are detected by classifiers trained in auxiliary databases by the usage of Transfer Learning (TL).

Initially, we have investigated how to deal with the dynamical information in BoVF representations for videos (Lopes et al. [2009d] and Lopes et al. [2009c]), using a *de facto* standard action database (Gorelick et al. [2007]) and different interest point detectors. In addition, we explored the usage of BoVF representations in other scenarios, like nude detection (Lopes et al. [2009a] and Lopes et al. [2009b]) and historical photographs classification (Batista et al. [2009b]). We also pursued a investigation on the cost-benefit of using the k-means algorithm against a random selection of visual words and a proposed enhancement over the pure random selection (Santos et al. [2010]).

We proceeded our investigation by addressing realistic action video databases, like those presented in Laptev et al. [2008] and Marszalek et al. [2009], collected from feature movies. As it will be seen in Chapter 4, contrarily to what happens in artificial, strictly controlled experiments, contextual information brings important clues for action recognition. In the proposed solution, the contextual information is expressed at high-level, in terms of concept occurrence probabilities, computed by BoVF-based SVM classifiers. In that solution, we also tackle the issue of lack of annotated databases

for training the Machine Learning (ML) algorithms involved, by assuming that the knowledge acquired from an external auxiliary database – for which a number of concepts have already been annotated – can be used to infer the context in the target action database. Such assumption is supported on the theory of Transfer Learning (Section 2.4).

1.1 Motivation

In recent years, Internet users witnessed the emergence of a great amount of multimedia content in the Web. Such kind of content is originally generated by professional or semi-professional individuals or enterprises, in a typical broadcast scheme. However, in a second wave, the users themselves start to create and publish their own multimedia productions. This motion towards user-generated content was stimulated by a number of factors, mainly the drop in cost of devices such as cameras and microphones, the spread of high-bandwidth connections and the emergence of Web 2.0 applications, including social networks. That new scenario led to an overwhelming increase in the amount of multimedia content available, which brought up the limitations of traditional Web tools in dealing with non-textual data.

One important step to make multimedia data effectively available is the high-level (i.e., semantic) indexing of such material, aimed at meaningful, semantically oriented retrieval. In the textual case, the words themselves convey quite directly the semantics. However, in the case of visual information, the connection between the low-level data in which they are encoded – ultimately, pixels – and the semantic meaning that human beings associate to them is far from immediate. Indeed, that is an open research issue, commonly referred to as the *semantic gap* (Smeulders et al. [2000]).

The current state-of-the-art approach in terms of systems for image and video retrieval is described in Snoek and Worring [2008][1]. Such systems are composed by several individual concept detectors which are applied independently to every item in the database. The computed probabilities of occurrence of each concept is then taken as components of the feature vectors which will represent those media items for the search engine.

Systems like that one just described have the advantage of enabling textual search, as opposed to query-by-example[2] or query-by-sketch[3] approaches, which are perceived

[1] Although that paper is focused on video retrieval, most of its ideas can also be applied for images.

[2] In query-by-example systems, the users choose an image as a query and the system returns those ones considered most similar to it.

[3] In query-by-sketch systems, the users draw a rough sketch of the image they want to find.

as unnatural for users accustomed to commercial search engines.

One key issue here is which concepts should be searched for. This issue was addressed in the Large-Scale Concept Ontology for Multimedia (LSCOM) workshop (Kennedy and Hauptmann [2006]), which defined a lexicon containing around 1000 concepts, from which 449 have been annotated in 80 hours of video coming from TREC Video Retrieval Evaluation (TRECVID) 2005 database (Over et al. [2005]).

Later on, experiments performed by Kennedy [2006] showed that annotations for some concepts defined in the LSCOM varied significantly whether the annotators watched video sequences or looked at keyframes. Those results indicate that the dynamic nature of video information plays an important role in the recognition of some concepts. Also, they suggested that such dynamics cannot adequately be captured by the direct application of techniques aimed at still images.

The 24 activity/event LSCOM concepts which had their annotations changed after the experiments described in Kennedy [2006] are listed in Table 1.1, from which it is possible to see that all those concepts, either directly or indirectly, are related to actions performed by human beings.

Table 1.1. LSCOM concepts re-annotated by using video segments instead of keyframes (Kennedy [2006]).

Airplane Crash	Greeting
Airplane Flying	Handshaking
Airplane Landing	Helicopter Hovering
Airplane Takeoff	People Crying
Car Crash	People Marching
Cheering	Riot
Dancing	Running
Demonstration Or Protest	Shooting
Election Campaign Debate	Singing
Election Campaign Greeting	Street Battle
Exiting Car	Throwing
Fighter Combat	Walking

1.1.1 Additional Applications

Although the improvement of high-level video indexing and retrieval is a major motivation for the present work, it is worth mentioning that action recognition techniques can be used in a variety of applications.

A great amount of work has been done around the idea of building "smart" video surveillance systems, which would be able to detect suspicious behavior automatically. In Lavee et al. [2007], for instance, a framework to aid the search for specific events in recorded surveillance video is proposed. In addition, recognition of people by their gait has been studied as alternative biometrics, as in Kale et al. [2004]. A review focused on visual surveillance systems similar to those is presented in Hu et al. [2004].

The analysis of sport videos is another widespread application. In Xie et al. [2004], for example, the classification of video segments between play and break intervals of soccer is suggested as a means of summarizing the video (by taking out the break intervals). Soccer games are also analyzed by Zhu et al. [2007a], in which text and the players' trajectories are used to build a system aimed at helping coaches in tactical analysis. Six actions of a cricket umpire are analyzed in Rahman and Ishikawa [2005] while Tong et al. [2006] proposed the usage of local motion analysis to identify different swimming styles.

Hand gestures can be analyzed both for recognizing sign language symbols – as in Cooper and Bowden [2007] and Buehler et al. [2009] – as well as for improving Human-Computer Interaction (HCI). For example, in the pioneer work of Bobick et al. [1999] an action recognition system was used to build the *KidsRoom*, an environment which is able to interpret and react to specific actions of a group of children in a room.

Hand gesture recognition was also used by Tan and Davis [2004] to identify segments in lecture videos that are worth transmitting in less compressed formats. The assumption here is that specific actions can indicate the importance of each sequence. In a similar application, Ren and Xu [2002] proposed a system called *smart classroom*, where the actions performed by a teacher are recognized to allow for automatic camera motion and virtual mouse.

A variation of the general content-based retrieval idea is what is called *intelligent fast forward* as proposed by Zelnik-Manor [2006], in which the query video segment is compared against other segments in the same video.

Action recognition is an important issue also in robotics, in which the interpretation of human actions can be used either for properly reacting to the recognized action (i.e., control) or for learning and imitation (Kruger et al. [2007]).

Finally, in the medical area, human motion analysis can aid the identification of

motor problems in patients by the comparison of their motion to normal patterns, like in Branzan Albu et al. [2007], for example. Another possible medical application is remote assistance of elderly people, as suggested in Yanik et al. [2011], in which visual and non-visual sensors are applied to monitor those people.

1.2 Objectives

A fundamental drawback of current approaches for human action recognition in realistic videos comes from the large intra-class variability expected in unconstrained scenarios. This becomes a problem because the currently most successful approaches are based on ML techniques, most of them demanding a great amount of training examples to be able to cope with that variability. By the other side, the manual collection and annotations of a large set of training examples from videos for a large spectrum of actions is an expensive, error-prone and yet-to-be-done task.

In this work, we tackle this issue by proposing and testing a methodology for action recognition aimed at realistic videos which adds contextual information to a Bag-of-Visual-Features (BoVF) video description and circumvents the lack of large annotated action databases by means of Transfer Learning (TL).

Before enunciating the main hypotheses of this thesis, a definition of *context* is needed:

Context

Context in a action video or image is any information about the environment where the action takes place.

As an analogy, if the region (i.e., the spatio-temporal volume) in which the action occurs would be segmented from the video, the ***action*** would compare to the *foreground* and the ***context*** would be compared to the *background* in a typical segmentation scenario.

Given the above definition of *context*, the proposed solution relies on the following hypotheses:

1. Contextual information (i.e., information coming from context) is relevant for action recognition.

2. Context can be described by high-level concepts appearing in the scene.

3. Transfer learning can be used to obtain accurate-enough concept detectors from auxiliary – already annotated – databases.

Additionally, it relies on the assumption that Bag-of-Visual-Features are adequate representations for realistic videos and images.

Those hypotheses and assumption are further detailed and justified in Chapter 4.

1.3 Contributions

The solution presented in previous section encompasses our main thesis, which was supported by the experiments and can be summarized in two main contributions:

- It is feasible to use TL techniques to detect concepts in realistic video action databases.
- By using the transferred information, it is possible to enhance action recognition in those scenarios.

The development of the main contributions led to some derivative ones:

- A study on the cost-benefit of the k-means algorithm for vocabulary building – Santos et al. [2010].
- A study on adding dynamic information to BoVFs (Lopes et al. [2009d]).
- Studies on applying BoVFs to other realistic, challenging contexts, namely, nudity detection in images – Lopes et al. [2009a] – and videos – Lopes et al. [2009b] – in addition to historical photographs categorization – Batista et al. [2009a] and Batista et al. [2009b].
- A Code base for research with BoVF representations (Section **??**).

Every one of these contributions are detailed in Chapter 5.

1.4 Thesis Organization

This text is split in two main parts: Part I presents the thesis fundamentals: firstly, methodological foundations for the developed approach are provided in Chapter 2. Next, a broad survey on different approaches for action recognition is presented in Chapter 3. That survey provided an organizing framework for the main kinds of

approaches to action recognition – Figure 3.1 – and guided the choice of an approach based on BoVF representations and SVM classifiers as the basis for our own solution.

Part II details the implemented solution in Chapter 4. The results achieved along the development of the thesis are then described in Chapter 5, along with citations of the papers in which they were published[4]. All papers published during the course are listed in Section 1.5.

Finally, some concluding remarks and pointers for future work are presented in Chapter 7.

1.5 Summary of Publications

This section provides a summary of the publications generated along the development of this thesis. Overall, they sum up to 13 published papers.

1.5.1 Publications on Human Actions

- Lopes, A. P. B., Santos, E. R. da S., do Valle Jr., E. de A., de Almeida, J. M., and Araújo, A. de A. (2011). Transfer Learning for Human Action Recognition. In Proceedings of SIBGRAPI '11.

- Santos, E. R. da S., Lopes, A. P. B., do Valle Jr., E. de A., de Almeida, J. M., and Araújo, A. de A. (2010). Vocabulários Visuais para Recuperação de Informação Multimídia. In Proceedings of WEBMEDIA '10.

- Lopes, A. P. B., Oliveira, R. S., de Almeida, J. M., and de Araújo, A. de A. (2009c). Comparing alternatives for capturing dynamic information in bag of visual features approaches applied to human actions recognition. In Proceedings of MMSP '09.

- Lopes, A. P. B., Oliveira, R. S., de Almeida, J. M., and Araújo, A. de A. (2009d). Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. In Proceedings of SIBGRAPI '09.

1.5.2 Publications on Related Applications

- Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., and Araújo, A. de A. (2009a). A bag-of-features approach based on hue-Scale-Invariant Fea-

[4]Portions of current thesis are verbatim or modified excerpts from those papers.

ture Transform (SIFT) descriptor for nude detection. In Proceedings of EUSIPCO '09.

- Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., Coelho, M. de M., and Araújo, A. de A. (2009b). Nude detection in video using bag-of-visual-features. In Proceedings of SIBGRAPI '09.

- Batista, N. C., Lopes, A. P. B., and Araújo, A. de A. (2009a). Detecção de edifícios em fotografias históricas utilizando vocabulários visuais. In Proceedings of CLEI '09.

- Batista, N. C., Lopes, A. P. B., and Araújo, A. de A. (2009b). Detecting buildings in historical photographs using bag-of-keypoints. In Proceedings of SIBGRAPI '09.

1.5.3 Publications on Other Applications

- Lopes, A., Flam, D. L., Batista, N. C., Avila, S. E. F., Almeida, J. M., and Araújo, A. A. (2008a). Automatic frame extraction for improving content-based image retrieval of historical photographs. In Proceedings of WEBMEDIA '08.

- Lopes, A., Oliveira, C., and Araújo, A. de A. (2008b). Face recognition aiding historical photographs indexing using a two-stage training scheme and an enhanced distance measure. In Proceedings of SIBGRAPI '08.

1.5.4 Publications from Collaborations

- Oliveira, J. E. E. d., Lopes, A. P. B., Chavez, G. C., Deserno, T., and Araújo, A. de A. (2009a). Mammosvd: A content-based image retrieval system using a reference database of mammographies. In Proceedings of the CBMS '09.

- Oliveira, J. E. E. d., Lopes, A. P. B., Chavez, G. C., Deserno, T., and Araújo, A. de A. (2009b). Mammosys: a content-based image retrieval system using breast density patterns. Computer Methods and Programs in Biomedicine, v. 99, p. 289-297

- Oliveira, R. S. and Ramos, T.L.A.S. and Lopes, A. P. B. and Oliveira, C. J. S. and Araújo, A. de A. Extensão do Algoritmo CCV Aplicado à Recuperação de Imagens com Base no Conteúdo. In: Workshop of Undergraduate Work(WUW) - SIBGRAPI '08.

Part I

Foundations

Chapter 2

Methodological Foundations

This chapter is aimed at establishing the foundations underpinning the current thesis. It starts by introduction of the concept of interest points in Section 2.1, which provides some detail on those specific algorithms that are used throughout this work. A detailed description of the building process of Bag-of-Visual-Features (BoVF) representations for visual data is provided in Section 2.2. Section 2.3 addresses the issue of classification with Support Vector Machines (SVM) and finally, in Section 2.5 presents some concluding remarks on the chapter.

2.1 Interest Points

Interest points (or keypoints, or local features[1]) are specific points (or regions) in an image or video which present some visual characteristics considered particularly important. Typically, algorithms of interest point detectors aim at characteristics such as efficiency, location accuracy, distinctiveness and invariance to a number of transformations on the images or videos.

Tuytelaars and Mikolajczyk [2008] point out the emergence of the powerful Scale-Invariant Feature Transform (SIFT) algorithm, from Lowe [2004], as a landmark for the recent resurgence of research based on interest points. SIFT was first presented in the context of finding distorted copies of identical objects in images and therefore has a strong emphasis on invariance. Indeed, the SIFT descriptor is invariant to scale, translation and rotation transformations, besides being robust to illumination changes and affine transformations.

[1] All those terms – *interest points*, *keypoints* and *local features* – are applied interchangeably by different authors.

The application of interest points to exact object recognition evolved to concept recognition approaches, most of them based on statistical analysis of interest points instead of exact matching of individual points. Such approaches – which, as it was seen in Chapter 3, are mostly BoVF-like – have proved more robust to a number of image transformations and occlusion, and thereby better suitable for image/video analysis at a higher level.

The continuous success of approaches based on interest points – and more specifically based on BoVF representations built on them – to static concept recognition paved the way to the application of the same ideas for action recognition. Additionally, it gave rise to a huge number of algorithms for interest point detection and description, as can be seen in survey papers like Tuytelaars and Mikolajczyk [2008] and comparative evaluation papers as Mikolajczyk and Uemura [2008] (for 2D descriptors) and Wang et al. [2009], for example.

In this section, we provide a brief presentation of the interest point detectors/descriptors used throughout this work.

2.1.1 Scale-Invariant Feature Transform (SIFT)

The SIFT (Lowe [2004]) is an algorithm for interest point detection and description which searches for points presenting invariance to scale, location and rotation, besides robustness to affine transformations as well as illumination changes. These characteristics turned SIFT interest points algorithm quite successful for object recognition and make it a natural candidate for extensions to video.

The SIFT detector achieves good efficiency by computing candidate points as the local maxima and minima in Difference of Gaussians (DoG) applied in the image space-scale. However, only the candidate keypoints which are stable are kept. Points with low contrast and on edges are also discarded. The scale in which the detection is more stable is selected as the scale for each point. Rotation invariance is achieved by computation of a canonical orientation determined by the peaks of a histogram of gradients computed on a gaussian window around the point. Finally, the descriptor is computed from 16 regions of 4×4 pixels around the point. A orientation histogram with eight bins and weighted by a gaussian centered on the point location is built in each region, providing a vector of $8 \times 16 = 128$ dimensions. SIFT algorithm is patented, but an executable code is available for research purposes[2].

[2] http://people.cs.ubc.ca/~lowe/keypoints/

The HueSIFT Descriptor

The HueSIFT descriptor is proposed in van de Sande et al. [2010] as a means to add color information to the SIFT descriptor. This is achieved simply by the concatenation of a hue histogram to SIFT. The hue histogram employed, by its time, weights the hue values by the saturation, to make the descriptor more robust van de Weijer et al. [2006] *apud* van de Sande et al. [2010].

2.1.2 Speed-Up Robust Features (SURF)

The SURF algorithm (Herbert Baya and Gool [2008]) pursues goals similar to the SIFT ones, but it is modified for better performance.

SURF detector search for points in which the determinant of the Hessian matrix has maximum values. In order to enhance performance, box filters which approximate second order Gaussian derivatives are used, since they can be computed at a low computational cost by using the concept of integral images of Viola and Jones [2001]. Instead of gray-level gradients, SURF descriptors are based on first order Haar wavelet responses. The concept of integral images is again applied for better performance and a smaller region around the point is considered, providing a vector with 64 dimensions instead of the 128 dimensions of SIFT. As in the case of SIFT, an executable code of the SURF algorithm is available for researchers[3].

SURF authors claim that it achieves results comparable to those provided by SIFT, but at a lower computational cost. Since the computational effort for processing videos is always potentially huge, we decided to empirically verify whether such claims hold in the specific setting of action recognition. The results of this comparison are presented in Section 6.2.

2.1.3 Space-Time Interest Points (STIP)

The STIP detector proposed by Laptev [2005] is an extension to the Harris corner detector to the 3D space. Such 3D space is composed by the spatial frames and the time dimension, as can be seen in Figure 2.1.

The original Harris detector finds points in which the gray-level values vary significantly in both directions (x and y) of an image. As in the case of SIFT, these points are obtained in the space-scale, defined by the convolution of the image with gaussian kernels with standard deviations σ_s, where s represents a specific scale. The proposed extension is aimed at finding points such as the image varies greatly both in space and

[3] http://www.vision.ee.ethz.ch/~surf/download.html

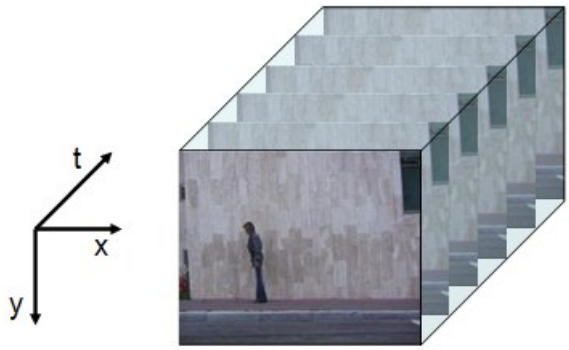


Figure 2.1. A video can be seen as a spatio-temporal volume composed of the spatial dimensions coming from the frames (y and y) along with the temporal dimension t (the image of the first frame is from the Weizmann database, presented in Gorelick et al. [2007]).

in time. Thus, the new scale-space is obtained by the convolution of the video with an anisotropic gaussian kernel defined not only by the variance in space (σ_s), but also by (an independent) variance in time (τ_s).

The original STIP detector as proposed in Laptev [2005] performed a scale selection procedure, by selecting only the points which presented both maximum values for the corner function and extreme values for the normalized spatio-temporal Laplace operator. Such procedure resulted in a extremely sparse set of points, which lead to poorer recognition results than denser samplings. Hence, the STIP version used in Laptev et al. [2008] and Marszalek et al. [2009] – the one which is publicly available – provide a multi-scale point selection, in order to obtain a denser set of points and also to reduce computational complexity, since the procedure for scale selection is eliminated from the algorithm.

The available STIP code[4] is also able to deliver two different types of descriptors (or both together, which is the *default* option): one based on Histograms of optical Flow (HoF) and another based on Histograms of Gradients (HoG) features. Both descriptors are computed in a 3D patch around the point, partitioned in $3 \times 3 \times 2$ blocks. HoG histograms are composed of 4 bins, while the HoF ones are composed of five bins. This provides a HoG descriptor with $(3 \times 3 \times 2) \times 4 = 72$ components and a HoF descriptor with $(3 \times 3 \times 2) \times 5 = 90$ components, which means the complete default STIP descriptor is composed of $72 + 90 = 162$ components.

[4] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

2.2 Bags-of-Visual-Features

Bag-of-Visual-Features are a visual analog to the traditional Bag-of-Words (BoW) representations for text retrieval (Baeza-Yates and Ribeiro-Neto [1999]). Figure 2.2(a) illustrates the concept of BoW representations for textual data. The main idea is to represent every text document as a histogram of word occurrences. This provides a compact representation for the texts which hopefully is able to keep their overall semantic content.

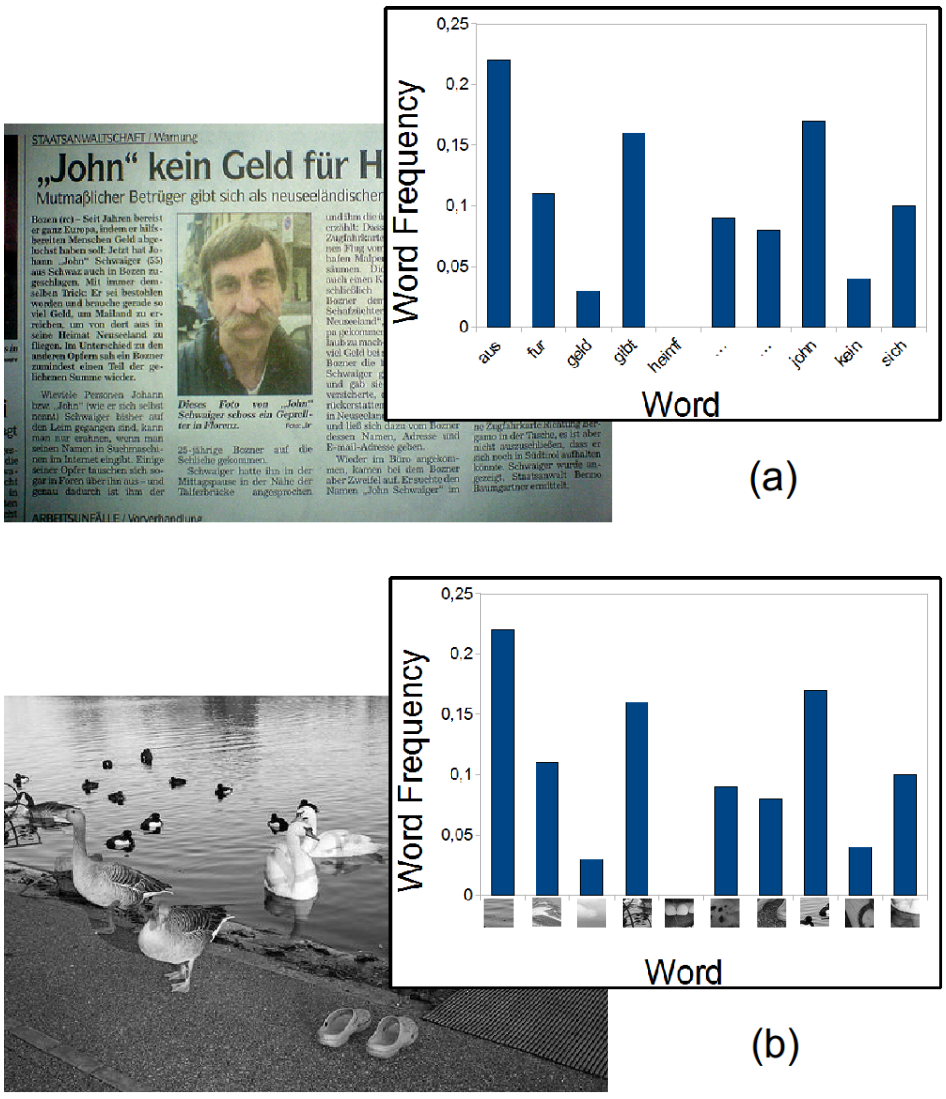


Figure 2.2. Schematics illustrating the relationship between: (a) representing text documents by BoWs, and (b) representing images by BoVFs. (Sources for the pictures: http://www.flickr.com/photos/ibcbulk/2473965083/ and http://www.flickr.com/photos/bbcworldservice/4403284481/ – both under Creative Commons licences: (a) Attribution-Noncommercial-Share Alike 2.0 Generic; (b) Attribution-Noncommercial 2.0 Generic).

Typically, the words included in the vocabulary (i.e., the bins of the histogram) are collected from a training *corpus* comprised of a large number of text documents. The simple computation of every different word form found in the training *corpus* typically provides a huge vocabulary size, leading to the so called *curse of dimensionality*. In order to alleviate this problem, some preprocessing is performed. First, words without semantic meaning like articles or prepositions (generically called *stop words*) are disregarded. Another common preprocessing is to group words by their roots, and considering each word family as a unique bin in the BoW histogram. Such grouping means, for example, that different conjugations of a verb are considered as the same *word* as well as gender and number variations for nouns.

Figure 2.2(b) illustrates the analogous idea for visual data. It is worth noticing that the only significant change from 2.2(a) to 2.2(b) is that in (b), the words representing histogram bins are replaced by image patches. This is an important difference between BoWs and BoVFs, because while in the textual case the definition of *word* is straightforward, the definition of a *visual word* is somewhat arbitrary, typically built from low-level data in terms of small patches. Next section provides the details of the general algorithm to build BoVF representations.

The general steps to define visual words and build BoVF representations for images or videos are depicted in Figure 2.3, which is detailed below.

2.2.1 Point Selection

The processing pipeline starts by selecting a set of points from the images or videos (step *a* in Figure 2.3). The simplest way to select such points is applying a 2D grid to the images or a 3D one to the videos spatio-temporal volumes. This is denominated *dense sampling* and leads to a huge computational complexity, both in space and time. To overcome this, more typical settings apply interest point detectors for a sparser selection of points. Nevertheless, as discussed in Dollar et al. [2005]; Niebles et al. [2008]; Mikolajczyk and Uemura [2008], for example, the best compromise between sparsity and informativeness is yet to be established. While in Laptev et al. [2008] the very sparse STIP interest point detector – which is highly focused in motion areas – is applied in a typical BoVF schema, in Mikolajczyk and Uemura [2008], for instance, several interest point detectors are applied with the explicit purpose of gathering a denser sample of points, carrying different kinds of meaningful information. In their case, though, the computational complexity of such a choice is explicitly addressed by parallel processing.

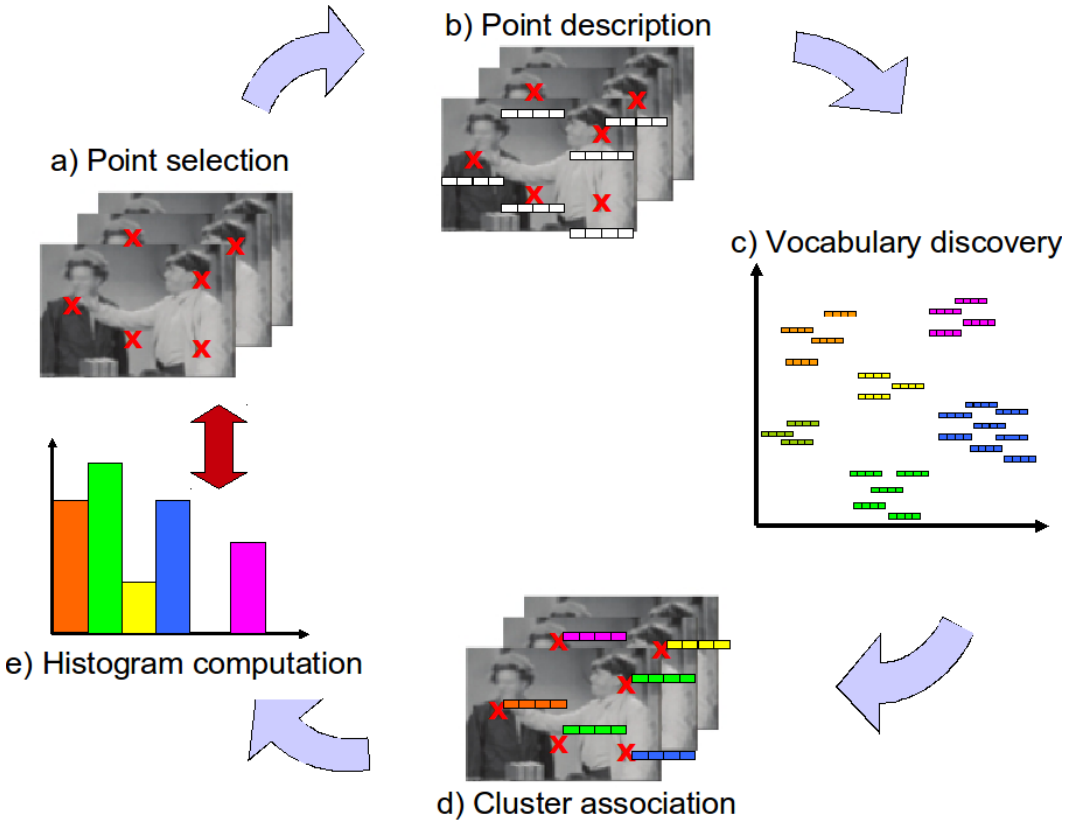


Figure 2.3. Steps for creating a Bag of Visual Features for visual data. The details of every step are provided in the text. Source: Lopes et al. [2009d].

2.2.2 Point Description

The step (b) in the BoVF-building pipeline is the visual description of the region around the selected points. Again, there is a variety of alternatives for this step, going from raw gray level values to those more sophisticated descriptors normally delivered by interest point algorithms. Typical interest point algorithms provide both the location of the points(detection) and at least one form of descriptors for those points, and using both detector and descriptor delivered by the same code is a straightforward choice. However, it is also possible to use descriptors which are distinct from those ones delivered by the chosen interest point algorithm. In the above mentioned work of Mikolajczyk and Uemura [2008], for instance, the same descriptor is computed from the interest points selected by all the detectors applied, while in Marszalek et al. [2009], 2D Harris corners are described by SIFT descriptors.

Algorithms for selection and description of interest point were previously addressed in Section 2.1.

2.2.3 Vocabulary Discovery

The potentially high number of local descriptors makes the direct representation of visual data by them computationally unfeasible in most cases. In addition, direct comparison produces matchings which are too specific. Although almost exact matchings are desirable for finding different instances of the same object, in the case of concept recognition – in which the concept instances come from different objects – direct matchings can be overly restrictive.

These issues are addressed by the quantization of the feature space. The usual approach is to use the k-means clustering algorithm (Mitchell [1997]) to perform such a quantization, by considering each cluster as *visual word*.

Nevertheless, most available descriptors are high-dimensional vectors, which can degrade both performance and quality of clustering results. To mitigate this problem, before clustering, the descriptors are typically submitted to some dimensionality reduction technique. The commonest choice for dimensionality reduction is the Principal Component Analysis (PCA) technique (Lay [2002]). Dimensionality reduction does not appear in Figure 2.3 because, although quite common, this is not a mandatory step to create a BoVF representation.

2.2.4 Cluster Association

After defining the quantized vocabulary, every descriptor on the videos is associated with one visual word (step d). This can be performed by computing their distances to the centroids of every cluster, which are learned in the preceding step. More sophisticated strategies for vocabulary learning, as that presented in Mikolajczyk and Uemura [2008], for example, can lead to different ways for associating descriptors to specific words.

2.2.5 Histogram Computation

The final step in BoVF computation is to count the occurrences for every visual word in every visual item (image or video) to form the histograms which will constitute the BoVF representation for those items (step e).

In order to deal with different number of local descriptors selected in each image or video, some form of normalization is performed. In a classical BoW, the words are weighted based on their relative occurrences both in the entire *corpus* and in every text, since too rare or too frequent words tend to have low discriminative capability. A classical approach for weighting words taking these facts into account is to compute

the term-frequency inverse-document-frequency (tf-idf) for the words in each document, instead of using a more simplistic normalized counting. Nevertheless, the experiments reported in Yang et al. [2007] suggest that in the visual case the tf-idf representation provides no significant improvement in image classification results. In other words, in the BoVF case, a simple normalization achieved by the computation of relative frequencies usually suffices.

* * *

The construction of the BoVF vectors corresponds to the representation step of Figure 3.1. The next stage for achieving a complete action recognition approach is to use these vectors for action modeling and classification. In this work, these too steps are performed by Support Vector Machines, which are detailed in next section.

2.3 Support Vector Machines

The recognition of human actions as proposed in this work is actually a classification problem, which generically can be posed as follows: given some *pattern* $\mathbf{x}$ representing an entity in the application domain and a set of possible (categorical) classes or labels $\mathcal{L}$, find the appropriate label related to that pattern[5].

This can be achieved by applying a function which maps every possible pattern to its corresponding label:

$$f : \mathbf{x} \mapsto y \in \mathcal{L} \tag{2.1}$$

Ideally, an analytical mapping function would be derived from domain knowledge and classification would turn into a matter of computing that function to every new pattern which needs to be classified. When such an analytical formulation can not be derived, one possible solution is to algorithmically try to *learn* the mapping function, using as hints a number of examples or *training patterns* for which one already knows the labels. In other words, looking at the training examples, the learning algorithm formulates a hypothesis h for the mapping function:

$$h : \mathbf{x} \mapsto y \in \mathcal{L} \tag{2.2}$$

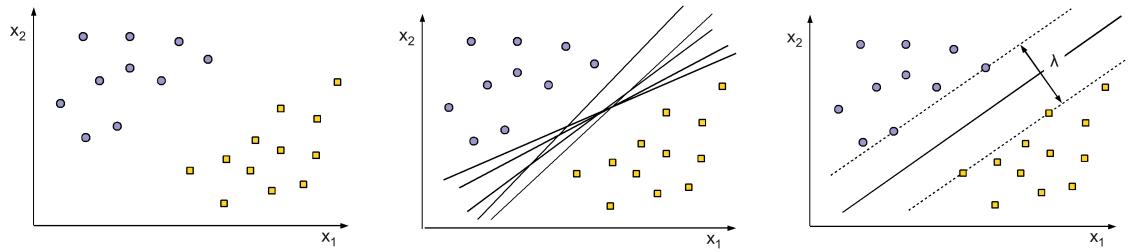
The goal of a learning algorithm is then to minimize error, defined as a function of the difference between the hypothesis function h and the actual mapping function f.

[5]The *pattern* terminology is applied after Scholkopf and Smola [2001], to stress the fact that there is no assumption about these patterns other than composing a set (i.e., they do not necessarily need to be vectors).

Until the 90's, the more successful algorithms for learning complex hypothesis functions were based on Neural Networks (NN) – Scholkopf and Smola [2001]. In the last years, though, neural networks have been gradually replaced by Support Vector Machines (SVM) classifiers as the off-the-shelf choice for complex classification problems. That is due to a number of advantages of SVM over NN, including a stronger theoretical foundation, the absence of false minima and a better generalization capacity. Additionally, SVM proved to be able to provide successful results in a great variety of practical applications, by the usage of appropriate kernels (see Section 2.3.2).

2.3.1 Linear Support Vectors Classifiers

Patterns are typically represented as vectors in some *input space*. Figure 2.4(a) illustrates a *binary classification* problem, i.e., when the label set is composed of only two classes – represented as $\mathcal{L} = \{-1, 1\}$ – and the patterns are bi-dimensional vectors $\vec{x} = (x_1, x_2)$, where x_1 and x_2 are features collected from the items to be classified. That figure shows a somewhat easy classification problem, which is said to be *linearly separable*. In such cases, to learn a classifier can be regarded as a problem of learning a *decision boundary* between the two classes, based on a number of *training vectors*.



(a) (b) (c)

Figure 2.4. (a) the binary classification problem for 2D patterns; (b) several possible linear decision boundaries; (c) the linear decision boundary which maximizes the margin λ between the classes.

In Figure 2.4(b), it is possible to see that an infinite number of straight lines – or hyperplanes, in case of input spaces with more than two dimensions – could be used to separate the same training data. A possible choice is the hyperplane which maximizes the distances from the closest vectors in each of its sides. The hyperplane defined by this criterium defines the *maximum margin classifier*. It is represented by the continuous line in Figure 2.4(c).

To find such a hyperplane is an optimization problem whose formulation is given by:

Maximize:

$$W(\alpha_i) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\left(\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \vec{x}_i.\vec{x}_j\right) \tag{2.3}$$

Subject to:

$$\alpha_i \geq 0 \tag{2.4}$$

and

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.5}$$

which is the dual formulation for the *hard margin classifier*[6]. The α_i values – the lagrange multipliers – are the values to be optimized in Equation 2.3. An important characteristic of Equation 2.3 that it is quadratic in the α_i values, which means that this is a convex optimization problem, and therefore has a unique global optimum.

The lagrange multipliers α_i can be computed by any generic routine for constrained quadratic optimization and then applied to the decision function which defines the class to which a testing point pertains, given by[7]:

$$D(\vec{x}_t) = sign\left[\sum_{i=1}^{m} \alpha_i y_i (\vec{x}_i.\vec{x}_t + b)\right] \tag{2.6}$$

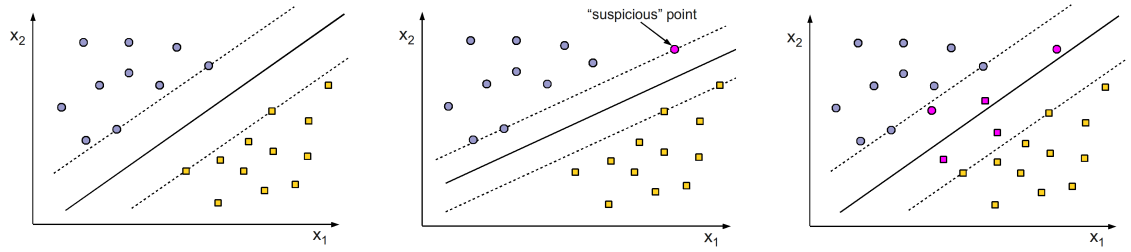
It is interesting to point out that the greater the lagrange multiplier α_i associated with some training pattern $\vec{x}_i$, the greater the influence of such pattern to the definition of the separating hyperplane. In fact, the training patterns which have $\alpha_i \neq 0$ are the only ones that define the margin and are called the *support vectors*. Since in most cases there are only a few support vectors, the classification time is significantly reduced when compared to NN.

Nevertheless, this also means that a "suspicious" point as the one emphasized in Figure 2.5(b) is going to produce a very large α value, therefore forcing the hyperplane towards a direction which can be quite different from that one which would be produced without taking that point into account (shown in Figure 2.5(a)).

Points like those just described can be either correct – yet unusual – examples, or outliers. In order to lessen the influence of a few noisy data, it is possible to introduce a relaxation on the hard-margin constraints by establishing an upper bound limit on

[6]The derivation of this formula is outside the scope of this text, but can be found in several books, including Scholkopf and Smola [2001].

[7]The value for the bias b is also defined by the α_i values.



(a) (b) (c)

Figure 2.5. Illustrating how an outlier can affect the definition of the separating hyperplane. The hyperplane in (a) is computed without the candidate outlier (marked as "suspicious" point), while in (b) the hyperplane computation takes that point into account. In the soft margin classifier (c), some points are allowed to be inside the margin or even at the wrong side of the separating hyperplane.

the α_i values:

$$0 \leq \alpha_i \leq C \tag{2.7}$$

In practice, such constraint means that some (potentially erroneous) points are now allowed to be inside the margin, as depicted in Figure 2.5(c).

Equations 2.3, 2.5 and 2.7 provide the formulation for the *soft-margin classifier*, best known as the *linear Support Vector Machine*. As in the hard-margin case, as soon as Equation 2.3 – subject to the constraints expressed by Equations 2.5 and 2.7 – is solved, the decision function can be computed to provide the class label for new patterns.

Next section describes how this formulation evolved to be able to cope with non-linearly separable data.

2.3.2 Classifing Non-Linearly Separable Data

Figure 2.6 presents some intuition on how initially non-linearly separable data can be made separable by the addition of more dimensions. In (a), it is not possible to draw any straight line which separates the two classes. However, by adding a third dimension and properly mapping the original points in it as done in (b), the classes become linearly separable in the new 3-dimensional space. More generically, the problem of classifying non-linearly separable data turns into a problem of finding out a function Φ which maps a pattern $\vec{x}$ from the original *input space* to a higher-dimensional *feature space* $\mathcal{H}$, in which, hopefully, the mapped data is separable:

$$\Phi : \vec{x} \mapsto \Phi(\vec{x}) \in \mathcal{H} \tag{2.8}$$

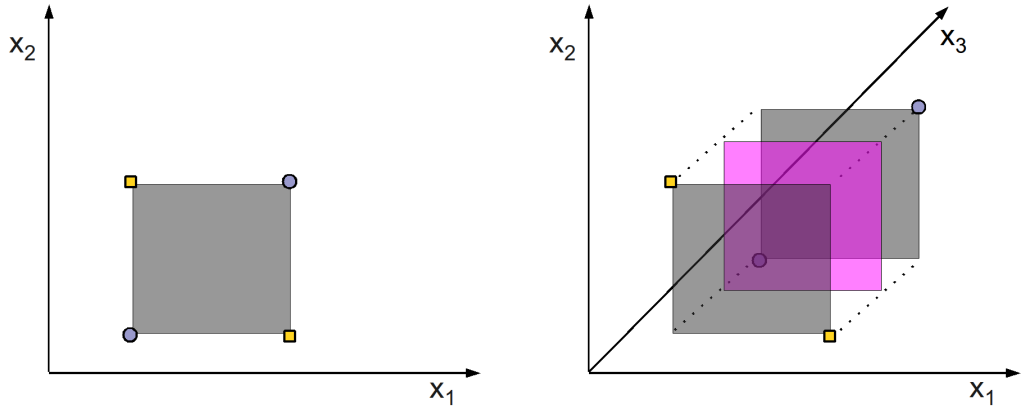


Figure 2.6. The addition of a third dimension along with a proper mapping can turn previously non-separable data (a) into linearly-separable data in the new 3D space (b). The magenta square between the two gray ones represents the separating plane in that new space (this picture was built based on the ideas verbally presented in Campbell [2008]).

The issue is how to find a proper mapping function Φ for every application domain. In most cases, an analytical formulation for this function simply cannot be found, because the relationship between input data and the classes is too complex. Additionally, even in cases when a suitable mapping can be found, its computation can be unfeasible in practice.

Nevertheless, one important property of the decision function expressed in Equation 2.6 is that it depends only on the dot product $\vec{x}_t.\vec{x}_j$. So, since the mapping function Φ maps $\vec{x}$ to a feature space $\mathcal{H}$ such as the dot product $\Phi(\vec{x}_i).\Phi(\vec{x}_j)$ is defined, the linear SVM can be applied in the feature space *without directly computing the mapping*. Actually, the dot product can be seen simply as a similarity measure, so virtually "any" function which is able to produce values which somehow reflect the similarity between $\vec{x}_i$ and $\vec{x}_j$ can be used as the dot product in the feature space. Such functions, which implicitly perform the mapping between the input space and the feature space, are the so called *kernels*:

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j) \tag{2.9}$$

To be a valid kernel, a candidate function needs to map the input space into the original patterns into a vectorial feature space in which the dot product is defined. Such spaces $\mathcal{H}$ are called *Hilbert spaces*. Kernel functions in Hilbert spaces have the property of satisfying the Mercer's conditions, whose mathematical formulation can be found in Scholkopf and Smola [2001], for example. However, Campbell [2008] point out that the Mercer's condition is translated into the constraint of the kernel function

be a positive semi-definite function. In practice, such condition can be verified by the analysis of the kernel matrix (also known as the Gram matrix), defined as:

$$K_m = \begin{bmatrix} K(\vec{x}_1, \vec{x}_1) & K(\vec{x}_1, \vec{x}_2) & \cdots & K(\vec{x}_1, \vec{x}_m) \\ K(\vec{x}_2, \vec{x}_1) & K(\vec{x}_2, \vec{x}_2) & \cdots & K(\vec{x}_2, \vec{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(\vec{x}_1, \vec{x}_m) & K(\vec{x}_2, \vec{x}_m) & \cdots & K(\vec{x}_m, \vec{x}_m) \end{bmatrix} \tag{2.10}$$

in which the $\vec{x}_i$ are the training patterns.

The kernel matrix K_m is positive-definite, if it is:

a) positive in the diagonal, i.e.:

$$K(\vec{x}_i, \vec{x}_j) \geq 0 \quad \forall \quad 1 \leq i, j \leq m \tag{2.11}$$

where m is the number of training samples; and

b) symmetrical, i.e.:

$$K(\vec{x}_i, \vec{x}_j) = K(\vec{x}_j, \vec{x}_i) \tag{2.12}$$

In other words, if the matrix K_m defined by Equation 2.10 – which is built on the actual training data – obeys the conditions expressed by Equations 2.11 and 2.12, the kernel can be considered valid for practical purposes.

The usage of kernel functions to apply the original linear SVM classifier on non-linearly separable data is known as the *kernel trick*. The kernel trick makes possible the widespread use of SVM classifiers in a number of fields. Later on, other learning algorithms are adapted to use the kernel trick, giving rise to an entire study field about kernel-based learning algorithms, as can be seen in Scholkopf and Smola [2001] and Cristianini [2004], for example.

2.3.3 Practical Considerations

The steps for using a SVM classifier can be summarized as follows:

1. Choose a suitable kernel.
2. Select error (C) and kernel parameters.
3. Train the classifier, by solving the correspondent optimization problem.
4. Use the learned decision function to classify new data.

In the next sections, some practical aspects of the above mentioned steps are considered.

The Choice of a Suitable Kernel

The lack of a principled approach to choose an appropriate kernel function for each domain is the main drawback of the SVM classifier. Nevertheless, good empirical results have been achieved using the most common kernels discussed in Section 2.3.2, namely the polynomial, the gaussian Radial-Basis Function (RBF) and the sigmoidal.

Campbell [2008] suggests that if there are few training points in a high-dimensional space, the data is very likely to be linearly separable already, meaning that the linear kernel should suffice (in that case, the suitability of the linear kernel can be checked out by verifying whether the training errors are near to zero). He also indicates the gaussian kernel as the first non-linear choice that should be considered.

Model Selection

Once a kernel has been chosen, the next step is often called the *model selection*, which is in fact a parameter tuning procedure. This is typically achieved by cross-validation. Ideally, this model selection via cross-validation should be performed on a separated sample, the *validation set*. This can become an important drawback for domains in which there is a limited amount of training data.

The theoretical reasons for the need of a validation set are discussed in Scholkopf and Smola [2001], the main one being the higher risk of overfitting. Nevertheless, they also point out that practitioners have been using the validation set as the final training set with excellent results. It is also common to "guess" suitable model parameters based on clues coming from previous similar experiences.

Finally, another practical advice provided by Scholkopf and Smola [2001] is to scale the kernels, to avoid numerical computation errors.

Training and Classification

The classifier is trained by the application of a quadratic constrained optimization technique, along with some "tricks" to deal with a large number of training points. After that, the classification is simply a matter of applying the learned classifier to new test data.

In fact, a number of mature SVM implementations are available through the internet, the LIBrary for Support Vector Machines (LIBSVM) library Chang and Lin [2001] being amongst the most used ones. LIBSVM is described in Section 2.3.5.

2.3.4 Kernels to Deal with Images

Multichannel Kernel

Despite of the flexibility allowed for the definition of kernel functions, it turned out that a few kernels proposed in the literature ended up successfully applied in many different domains. Among the most widespread kernels are the *polynomial kernel*, the *Radial-Basis Function (RBF) kernel* – more specifically the *gaussian RBF kernel* – and the *sigmoidal kernel* (Cristianini [2004] and Scholkopf and Smola [2001]).

The RBF kernel is generically expressed by:

$$K(\vec{x}_i, \vec{x}_j) = f(d(\vec{x}_i, \vec{x}_j)) \tag{2.13}$$

where $d(\vec{x}_i, \vec{x}_j)$ is a metric on the input space and f is a function on $\mathbb{R}_0^+$.

A particular and very common kind of RBF kernel is the gaussian kernel:

$$K(\vec{x}_i, \vec{x}_j) = exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right) \tag{2.14}$$

in which σ is the standard deviation.

In Chapelle et al. [1999] more generic version of the gaussian RBF kernel is proposed for image classification based on color histograms. The expression for that kernel is:

$$K(\vec{x}_i, \vec{x}_j) = exp\left[-\frac{1}{A}D(\vec{x}_i, \vec{x}_j)\right] \tag{2.15}$$

where $D(\vec{x}_i, \vec{x}_j)$ is a distance measure between $\vec{x}_i$ and $\vec{x}_j$ and A is a scaling parameter.

In Zhang et al. [2007a], the kernel in Equation 2.15 is applied to BoVF representations aimed at image classification, this time using information coming from multiple channels:

$$K(\vec{x}_i, \vec{x}_j) = exp\left[-\sum_{c=1}^{n}\frac{1}{A_c}D_c(\vec{x}_i, \vec{x}_j)\right] \tag{2.16}$$

in which n is the number of channels.

The multi-channel kernel of Equation 2.16 is used for action recognition both in Laptev et al. [2008] and Marszalek et al. [2009]. In those cases, A_c takes the value of the average distance for every channel [8] and the χ^2 distance is applied, as follows:

$$\chi^2(\vec{x}_i, \vec{x}_j) = \frac{1}{2}\sum_{k=1}^{d}\frac{(x_{i,k} - x_{j,k})^2}{x_{i,k} + x_{j,k}} \tag{2.17}$$

[8] This is a good example of what Scholkopf and Smola [2001] would call an "educated guess" (see Section 2.3.3).

in which d is the dimension of the input space (i.e., the number of features) and $x_{i,k}$ is the k_{th} feature (component) of the vector $\vec{x}_i$.

The multi-channel kernel makes it straightforward to compare the effects of different channels in the well-established gaussian kernel framework. Furthermore, by using the same kernels of Laptev et al. [2008] and Marszalek et al. [2009] we aim at keeping our results better comparable to theirs.

Pyramid-Matching Kernel

The Pyramid-Matching kernel is proposed in Lazebnik et al. [2006] aiming at adding some structural information to a classical BoVF. In it, the image space is split into several levels (or resolutions), meaning that the image is split l times in each dimension at level l. The corresponding parts are compared by a histogram intersection:

$$I^l = \sum_{i=1}^{L} min(H_x^l(i), H_y^l(i)) \tag{2.18}$$

Equation 2.18 shows the definition of the histogram intersection I^l at a given level l[9].

In the final kernel, each I^l is weighted according to the level:

$$w_l = \frac{1}{2^{L-l}} \tag{2.19}$$

where L is the total of levels taken into account for kernel computation.

This weighting scheme diminishes the importance of matches in larger cells, since they have a greater probability of being matches between dissimilar features.

Putting together Equations 2.18 and 2.19 and after some algebraic manipulation, the pyramid-matching kernel is given by:

$$k = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{2.20}$$

where I^l is the histogram intersection of the two images being compared.

[9] In this case, the histogram is the BoVF for each part of that level.

2.3.5 The LIBrary for Support Vector Machines (LIBSVM)

LIBSVM[10] Chang and Lin [2001] is a freely available library which implements both the classical C-SVM (presented in this text) and the alternative ν-SVM formulations[11]. After training, if a test data set with known labels is provided, LIBSVM outputs not only the predicted labels, but also the classification accuracy, given by:

$$accuracy = \frac{number\ of\ correctly\ predicted\ test\ examples}{total\ of\ examples} \times 100\% \tag{2.21}$$

A number of proposals have been presented to use the binary SVM classifier for multiclass classification (see Scholkopf and Smola [2001], chapter 7, for example). In LIBSVM, the *one-against-one* approach is implemented. In other words, binary classifiers are trained for every combination of two classes. Then, a voting based on the results of all classifiers is used to decide the class for each test example. According to the LIBSVM author, the choice of this approach for multi-class classification is motivated by the experimental comparison presented in Hsu and Lin [2002].

The usage of alternative kernels in LIBSVM is straightforwardly performed by the computation of the kernel matrix as defined in Equation 2.10, which is then passed as an argument to the training procedure.

Another characteristic of LIBSVM which is fundamental for this work is its capability for computing the probability estimates for each possible label. There are a number of approaches for estimating such probabilities, some of them presented in Wu et al. [2004], which also proposes the approach actually implemented in LIBSVM.

2.4 Transfer Learning

Classical Machine Learning (ML) algorithms rely on the assumption that training and test data come from the same probability distribution. In fact, though, it is rare a case of practical application in which such an assumption is concretely guaranteed. Most applications of ML randomly splits the available data between training and test sets (or among validation, training and testing sets) and assume that future real data to which the trained algorithm will be applied will follow the same distribution.

Nevertheless, as it is argued in Pan and Yang [2009], there are plenty of real-world examples in which such assumption is not realistic. A typical example is the

[10] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[11] The ν-SVM formulation uses a different parameter ν to control the fraction of training errors and of support vectors to be used.

classification of web pages, in which data is constantly changing, thus letting the underlying trained algorithms outdated. They go on by citing sentiment classification of products in shopping websites. Such classification is based on user review data, which can present huge variations among different categories of products. In this case, a sentiment classification algorithm trained on a few products will probably fail on most other products, and even specialized classifiers will not be able to cope with the variations of perceptions of the product users along time. In some other cases, as in ours, there is not enough available labeled data in the target database, and its acquisition is too expensive and error-prone.

Transfer Learning techniques come up to deal with those kinds of applications, by allowing distributions, tasks and domains of training and test data to vary. The notation, definitions and classification of Transfer Learning (TL) algorithms of Pan and Yang [2009] are going to be applied in most of this review of the topic.

According to them, a *domain* $\mathcal{D}$ is composed of a *feature space* $\mathcal{X}$ and a *probability distribution* over that space $\mathcal{P}(\mathcal{X})$, and can be mathematically defined by:

$$\mathcal{D} = \{\mathcal{X}, \mathcal{P}(\mathcal{X})\} \tag{2.22}$$

A *task* $\mathcal{T}$ is defined by a *label space* $\mathcal{Y}$ and a *prediction function* $f(.)$ which is the equivalent to the conditional probability $\mathcal{P}(\mathcal{Y}|\mathcal{X})$:

$$\mathcal{T} = \{\mathcal{Y}, \mathcal{P}(\mathcal{Y}|\mathcal{X})\} \tag{2.23}$$

The target domain $\mathcal{D}_t$ is the domain of interest for the application, usually with few or no labeled data at all. A source domain $\mathcal{D}_s$ is an auxiliary source of information, generally – but not always – with a great amount of labeled data, which hopefully can be used to aid the ML algorithm to be adequately trained for the target task.

From those definitions, three main categories of TL algorithms can be identified: inductive transfer learning, transductive transfer learning and unsupervised transfer learning.

Inductive Transfer Learning

Inductive Transfer Learning (ITL) concerns those cases in which the source and target tasks differ, no matter their domains. The most common case of ITL is when there is a lot of labeled data in the source domain, which makes this type of ITL similar to multitasking learning. For example, Dai et al. [2007] proposes the TrAdaBoost

extension of the AdaBoost algorithm to deal with new instances coming in an already trained system, potentially provoking a continuous distribution change.

A less common case of ITL occurs when there is a related source domain, but their instances are all unlabeled. An example of such a technique – more known as self-taught learning – can be found in Raina et al. [2007] in which a image representation is learned from unlabeled instances, to be applied in the target domain (whose task is also unsupervised).

Transductive Transfer Learning

Transductive Transfer Learning (TTL) involves cases in which the tasks are the same, but the domains vary. Looking at Equation 2.22, it is possible to see that different domains can vary in two aspects: they can have different feature spaces (for example, text classification for different languages) or they can share the same feature space but have varying probability distributions (for example, text classification on different specialized databases).

Unsupervised Transfer Learning

Unsupervised Transfer Learning (UTL) techniques are developed to the cases in which there is no labeled data on neither source or target domains. This type of transfer was used in our tests on vocabularies coming from the source database applied to describe the target database items (Section 5.1).

2.4.1 Sources of Knowledge in TL

In all cases of TL, the knowledge can be extracted from the source instances, from learned feature representations or from model parameters. In case of relational data-sources, it is possible to occur the transfer of relational knowledge either.

This can be seen in Davis and Domingos [2009], which provides a coarser classification between shallow transfer and deep transfer. In shallow transfer, test instances come from the same application domain but they are known to have a different probability distribution. In deep transfer, test instances comes from a different, mostly unrelated domain. In their case, an algorithm called Deep Transfer via Markov Logic (DTM) is created to transfer knowledge among classification of yeast protein genome, web pages labeling and social network related tasks. In other words, what Davis and Domingos

[2009] calls deep transfer is similar to transfer of relational knowledge as described by Pan and Yang [2009].

2.5 Concluding Remarks

This chapter presented the main concepts and techniques used in this work. Section 2.1 presents the concept of interest points and briefly described the specific algorithms included in the proposed experiments. Section 2.2 details the steps for building a classical BoVF representation for visual data and discusses some significant issues that arise in each one of those steps. In Section 2.3 it is pointed out that action recognition is indeed a classification problem and then presents the Support-Vector Machines classifier, the current *off-the-shelf* classifier applied by most BoVF-based approaches for action recognition.

Going back to Figure 3.1, interest points and BoVF-building are related to the representation extraction step of action (or concept) recognition, while the action modeling step is performed in the SVM training phase and action recognition step is equivalent to the SVM classification phase.

Finally, Transfer Learning (TL) theory is described, as proposed in Pan and Yang [2009] and Davis and Domingos [2009] and, using their concepts, our work is defined as an application of shallow inductive transfer.

Chapter 3

State of the Art

This chapter presents a survey of human action recognition approaches based on visual data recorded from a single video camera. We propose an organizing framework which puts in evidence the evolution of the area, with techniques moving from heavily constrained motion capture scenarios towards more challenging, realistic, "*in the wild*" videos. The proposed organization is based on the representation used as input for the recognition task, emphasizing the hypothesis assumed and thus, the constraints imposed on the type of video that each technique is able to address. Expliciting the hypothesis and constraints makes the framework particularly useful for selecting a method, given an application. Another advantage of the proposed organization is that it allows categorizing newest approaches seamlessly with traditional ones, while providing an insightful perspective of the evolution of the action recognition task up to now.

3.1 Related Surveys

An extensive survey on earlier studies about motion-based recognition was first presented in Cedras and Shah [1995]. For the authors of that work, the first step in motion-based recognition is the extraction of motion information from a sequence of images, which can be done by optical flow or motion correspondence. Motion correspondence is established by tracking specific points of interest through frames, and generating motion trajectories, which can be parameterized in several ways. Instead of computing motion information from the entire image or from specific points, region-based motion features can also be extracted. Explicit human body models are used to guide the tracking step.

The survey presented in Aggarwal and Cai [1997] is devoted to human motion

analysis which, for them, comprises the following overall steps: a) segmentation; b) joint detection; and c) identification and recovery of 3D structures from 2D projections. The authors characterize body structure analysis as either model-based or non-model-based, depending on whether or not an *a priori* shape model is used. The *a priori* models considered can be stick figures, contours or spatio-temporal volumes. The proposals reviewed in Aggarwal and Cai [1997] are also split into two broad categories – more related to action modeling and recognition steps (see Figure 3.1 in Section 3.2) – which are: *space-state models*, in which each static posture is considered as a state and state transitions occur with certain probabilities; and *template matching models*, where a template is computed for each action, and then a nearest-neighbor classification scheme is applied to recognize similar actions.

While the authors in Aggarwal and Cai [1997] focus on approaches based on models of the human body, a survey specifically focused on models for recognition of hand gestures is presented in Pavlovic et al. [1997].

By their turn, Gavrila [1999] reviews approaches modeling either the whole body or the hand. Selected papers are organized into three categories: *2D approaches without explicit shape models*, *2D approaches with explicit shape models* and *3D approaches*. This survey also provides (in its Table 1) a comprehensive list of applications envisioned at the time, organized into five groups: virtual reality, smart surveillance systems, advanced user interfaces, motion analysis and model-based coding. It is worth noticing that most of the prospective applications suggested by Gavrila [1999] still remain as unsolved challenges.

In Shah [2003], the recognition of human action is described as comprising the following – more general, in comparison to Aggarwal and Cai [1997] – steps: a) extraction of relevant visual information; b) representation of that information in a suitable form; c) interpretation. The specific modeling of human body or body parts is not seen by the author as an essential step for human action recognition. In contrast, tracking and trajectory computation are considered the primary subtasks. Therefore, this survey is focused specifically on trajectory-based techniques.

For Wang et al. [2003], human motion analysis comprises the following steps: a) motion segmentation; b) object classification of segmented moving regions – which can be shape or motion-based; c) tracking of identified objects along consecutive frames; and d) recognition of motion patterns, providing what they call *behavior understanding*. As in Aggarwal and Cai [1997], action modeling approaches are distinguished between *template-based* and *space-states based*.

The authors of Buxton [2003] present a survey from the perspective of the generative learning algorithms applied to any of the various processing steps of action

understanding systems. In that paper, such systems are categorized generically into *explicit models* and *exemplar-based models*.

In Wang and Singh [2003], the focus falls again onto approaches relying on human-body or body parts. That survey is mainly motivated by biometrics applications, and the paper is composed of two main parts: in the first, the author provides a detailed survey on tracking techniques applied to heads, hands or the whole body. In the second, techniques for analyzing different models for those tracked elements are reviewed.

The work of Aggarwal and Park [2004] expands and updates the earlier review presented in Aggarwal and Cai [1997], by including not only actions, but also interactions. The surveyed approaches are distinguished by the level of detail in which the moving objects are described. *Coarse level approaches* are those in which people are considered as bounding boxes or ellipses. Then, motion patterns are used to model the actions. In approaches lying in the *intermediate level* of detail, people are represented by large body parts or silhouettes. Finally, *detailed models* can be built on the entire body or on specific parts, such as hands in the case of gesture recognition tasks.

Action modeling approaches are also distinguished based on two different aspects: the first differentiates *direct recognition* from reconstruction of *body models* before recognition. The second aspect distinguishes approaches by their *static* or *dynamic nature*, if the recognition is performed on a frame-by-frame basis or taking the entire sequence as the basic unit analysis. High-level recognition schemes – similar to those which Wang et al. [2003] call behavior understanding – are also discussed, most of them relying on manually constructed semantic models of the world.

In Hu et al. [2004] an extensive review on papers related to surveillance systems is provided. The authors consider a visual surveillance system comprising of the following steps: a) motion detection, which includes object modeling, segmentation and classification; b) tracking of moving objects; and c) behavior understanding. For some applications, an additional step of natural language description can be added. The ability to identify people at a distance (gait-based recognition) can also be introduced.

In another review focused on trajectory-based approaches, Chellappa et al. [2005] define what it is called *activity inference*, comprised, in their view, of three steps: a) low-level video processing; b) trajectory modeling; c) similarity computation. In this scheme, low-level processing is aimed at computing trajectories for selected objects. The trajectories for each action can be modeled by varied techniques and for each model a similarity measure needs to be established.

The survey of Moeslund et al. [2006] offers an overview of human motion analysis in general, with a section devoted to action recognition. They suggest that action recognition approaches can be broadly separated between the ones that explicitly con-

sider human presence in the scene and the ones that do not. The recognition section of that paper is structured around three different kinds of tasks: *scene interpretation*, without identifying particular objects; *holistic recognition*, using the human body or body parts, to recognize both the subjects and the actions performed by them; *action primitives and "grammers"*, in which motor primitives are used for representation or control. The primitives in the latter task can be used to create an action hierarchy that gives a semantic description of the scene. However, in most of such approaches motion primitives are usually taken as already available.

The review of Kruger et al. [2007] is focused on robotics applications, more specifically for learning and imitation. They distinguish approaches based on: *scene interpretation*, in which the moving objects are not "identified", but have only their overall motions analyzed; the *body as a whole*; *body parts* and *grammars*.

The review presented in Poppe [2007] deals specifically with pose estimation, assumed as a needed step for action recognition.

In the recent short review presented in Ahad et al. [2008], which is explicitly focused on papers from 2001 to 2008, a hierarchical terminology composed of *action primitives*, *actions* and *activities* is adopted. This survey categorizes different proposals according to the ML techniques applied, regardless of the underlying representation. In other words, it is focused on the modeling step (Figure 3.1).

The major branches presented in Turaga et al. [2008] differentiate between approaches aimed at *actions* and those aimed at *activities* recognition. In their case, similarly to Moeslund et al. [2006], actions are defined as simple motion patterns executed by a unique human, while activities are more complex patterns, normally involving more than one person. The following four major steps for action recognition are identified: a) collecting input video; b) extracting low-level features; c) extracting mid-level action descriptions; and d) high-level semantic interpretations.

The low-level features considered in that survey are *optical flow*, *point trajectories*, *blobs and shapes separated from the background* and *filter responses*. According to them, actions can be described at mid-level by *non-parametric*, *volumetric* and *parametric* models. Actions and activities can be modeled either by *graphical models*, *syntactic grammars-like approaches* or *knowledge/logic based approaches*.

The work of Ren et al. [2009] reviews motion recognition approaches in the context of Content-Based Video Retrieval (CBVR). Two major approaches are identified. In *trajectory-to-trajectory approaches*, motion trajectories are extracted and compared for recognition; the category of approaches that take into account the internal structure of the object over time are denominated *sequence-to-sequence approaches*.

In Poppe [2010], only papers aimed at recognizing full body actions are taken

into account. Image representations are separated into three large groups: *global*, when a specific Region of Interest (ROI) is described globally, *local*, based either on interest points and densely sampled ones, and *application specific*. In his view, action classification can be performed either by *direct classification*, using the information coming from all the frames in the sequence together and *temporal state-space models*, in which action sequences are broken in smaller steps.

3.2 Categorizing Different Approaches for Action Recognition

Regardless of the application envisioned, the process of recognizing human actions from videos can be seen as comprising the three major steps, as depicted in Figure 3.1.

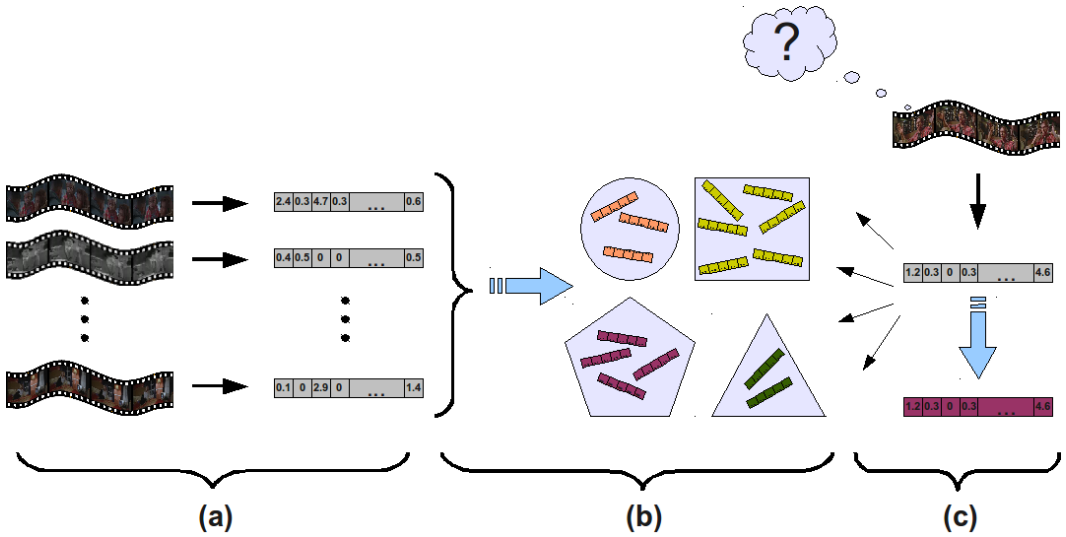


Figure 3.1. Overview of the processing steps needed for action recognition in videos: (a) *representation extraction* step, (b) *action modeling* step, and (c) *action recognition* step (picture best viewed in color).

(a) ***Representation Extraction***: this step starts with the extraction of low-level features from the videos, like color, texture and optical flow, for example. Those features are usually fed to a somewhat complex processing chain until a suitable (i.e., compact and descriptive) representation is achieved. It is worth noticing that, unlike Turaga et al. [2008], for instance, the output of this step is the final video representation which is used as the input to the action modeling step (below), regardless of the abstraction level. This generic definition is then applied to the finer-grained hierarchical structure proposed later in this section.

(b) ***Action Modeling***: in this step, the representations built in the previous step are mapped into different action categories. The *spectrum* of modeling alternatives goes from the selection of a small number of action templates aimed at direct comparison to sophisticated modeling schemes involving ML techniques.

(c) ***Action Recognition***: this last step takes place when unlabeled (query) videos are analyzed against the previously built action models, so that those videos can be associated with one of the possible action categories (i.e., classified).

As expected, those three steps are tightly interconnected. Some representation choices are more suitable for – or are even designed specifically to go with – certain kinds of techniques for action modeling. In the same way, the selected action modeling technique will determine – in some cases, in a unique way – how the classification is going to be performed.

The structure proposed in this section for organizing different approaches for action recognition is based on the representation step depicted in Figure 3.1. Such choice is justified by the fact that the process of extracting a specific representation for videos is directly related to a number of assumptions about the scene content. Such assumptions, by their turn, impose specific constraints on the types of videos that each recognition approach is able to cope with. Hence, an organization of different approaches which is built based on the selected representation provides a better ground to understand the strengths and limitations of each category of approaches, making it easier to: a) select appropriate approaches for specific applications, and b) distinguish which approaches are truly comparable among them. Finally, the selected *criteria* based on the underlying representation allow for sensibly unraveling unrelated approaches that end up mixed together under other categorization schemes.

Regardless of all the existing surveys discussed in Section 3.1, authors of new approaches follow no standard categorization structure while referring to previous papers that are related to their own approaches. For instance, in Ebadollahi et al. [2006b], authors propose a broad categorization between *object centric* and *statistical* approaches, while in Laptev and Perez [2007], authors distinguish among approaches based on *3D tracking of different points* of human bodies, *accurate background subtraction*, *motion descriptors on regions of actions* and *learning of actions models*. In Ke et al. [2007a], human action approaches are categorized into those based on *tracking*, *flow*, *spatio-temporal shapes* and *interest points*. In Zhao and Elgammal [2008], different approaches are categorized as *model-based*, *spatio-temporal template-based* and *bags-of-visual-features-based*. In the work of Wang et al. [2008], previous papers are coarsely

classified according to their specific goals, distinguishing among approaches dealing with *unusual event detection*, *action classification* and *event recognition*.

The categorization scheme we propose in this paper is depicted in Figure 3.2. In a coarse level, the different approaches are split into two large groups, which are nearly equivalent to the object centric *versus* statistical categorization proposed by Ebadollahi et al. [2006b]. It can also be considered a generalization of the categorization of Moeslund et al. [2006] into approaches that either consider the presence of humans or not. In fact, the framework proposed in this paper is a refinement of the model-based *versus* model-free categorization presented in Lopes et al. [2009b], although the terms *model-based* and *model-free* are abandoned in order to avoid confusion between action modeling and object modeling, the former being always present (Figure 3.1). The proposed scheme stresses the distinction between approaches that explicitly assume the presence of moving objects under specific conditions – like, for example, homogeneous background – from those in which such explicit assumption is not found.

As it will be seen, there is a non-negligible correlation between the proposed taxonomy and the temporal evolution of approaches: more recent approaches tend to rely on less constrained assumptions and therefore, more general hypothesis.

The two initial categories are further refined into subcategories organized according to the underlying representation. The vast majority of proposed solutions to human action recognition to date lies in the first broad group of approaches depicted in Figure 3.2. In other words, the video representation used by them explicitly assumes that one or more moving objects appear in the scene, typically under a number of specific conditions, like stable backgrounds and fixed scales, for example.

The basic idea behind those approaches is that it is possible to infer the actions being performed by studying the structure and/or the dynamics of the moving objects in the scene (or their parts). Moving objects of interest can be the human body, some body parts or other objects related to the application domain, like airplanes and automobiles, for example. Unlabeled moving regions can also be considered. In order to be able to analyze the moving objects, they need to be detected (and often, also tracked) before any further processing. Once the object has been detected/tracked, it can be either a) adjusted to some pre-defined model of the object, characterized by a number of parameters (parameterized object models), or b) characterized by global descriptors computed on their segmented area (implicit object models). Approaches relying on the presence of specific moving objects in the scene are further discussed in Section 3.3.

More recently, a number of approaches that do not explicitly rely on the presence of any specific object in the scene have been proposed. They are based on global statis-

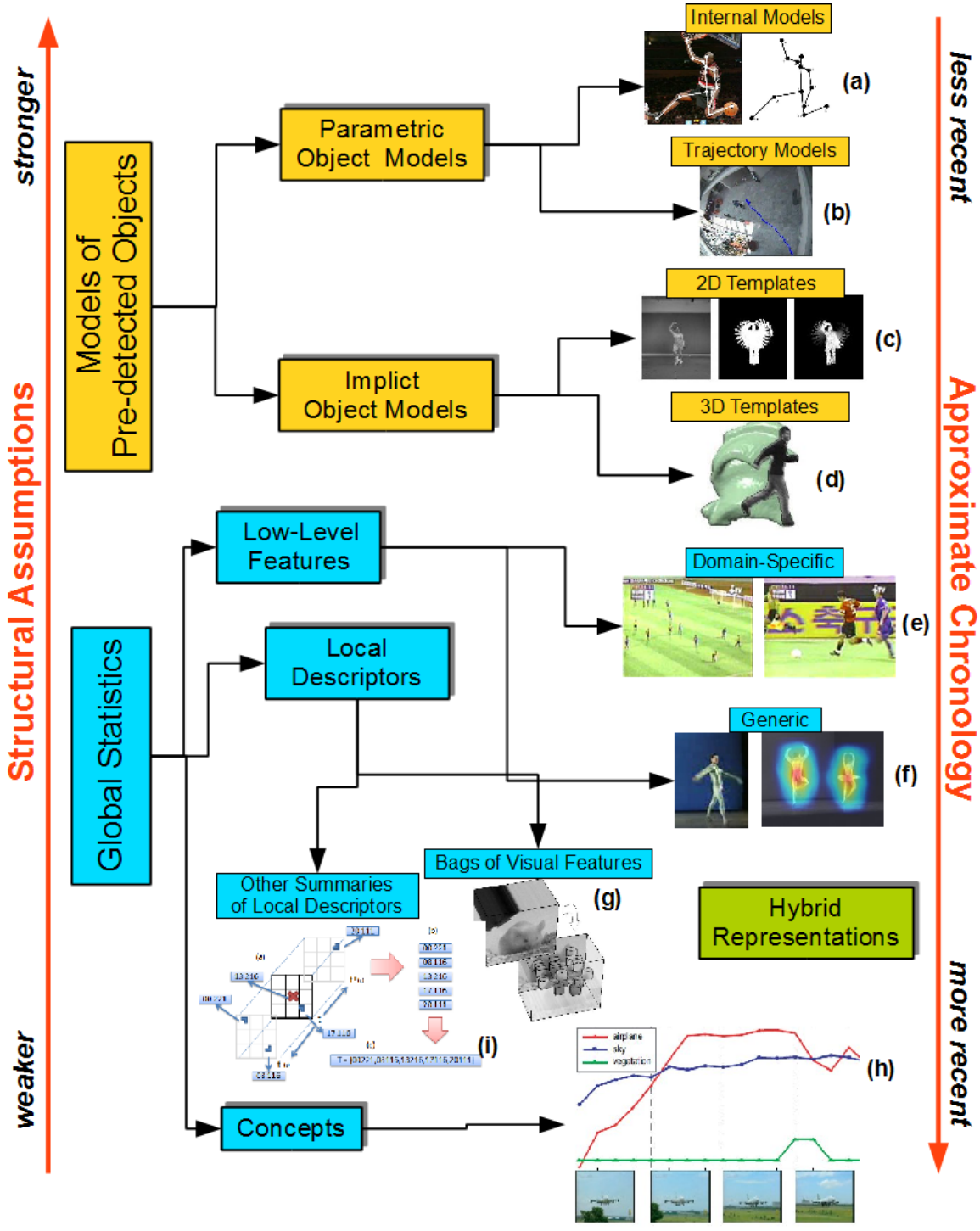


Figure 3.2. Categorization framework used along this survey for organizing the approaches for human action recognition found in the literature. It is based on the underlying video representations. Image references: (a) Yilmaz and Shah [2005], (b) Datong et al. [2008], (c) Bobick and Davis [2001], (d) Blank et al. [2005], (e) Xie et al. [2004], (f) Shechtman and Irani [2005], (g) Dollar et al. [2005], (h) Ebadollahi et al. [2006a], (i) Gilbert et al. [2008] (picture best viewed in color).

tical computations over different kinds of representations, within distinct abstraction levels: low-level features, mid-level interest points and high-level concepts. Approaches based on global statistics are represented by the lowest large rectangle in Figure 3.2 (which starts the branch in turquoise). As such approaches have only recently been proposed, previous surveys, discussed in Section 3.1, do not cover them in much detail. We provide a more comprehensive discussion of those approaches in Section 3.4.

Finally, hybrid approaches, which mix ideas from both those ones based on models of pre-detected objects and those ones based on global statistics can also be found.

Tables 3.1, 3.2 and 3.3 summarize the approaches for action recognition discussed throughout this chapter, pointing out the main assumptions for each category, and citing related papers that are going to be discussed in the following sections.

Table 3.1. Summary of Action Recognition Approaches Based on Object Model Representations (table best viewed in color).

Basic Representation – Object Models *Main Assumption: Actions can be derived from a specific model of the related objects.* [Related Papers by Subcategory (below)]			
Parametric Models *Objects related to actions obey a predefined model.*		Internal Models *Global representations of objects internal areas implicitly define the object model.*	
Internal Models *A predefined object model must describe objects internal states.*	Trajectories *The relevant information is not in the objects internal state, but in their positions over time.*	2D Descriptors *Objects appearance information is enough for action recognition OR motion can be aggregated in 2D representations.*	3D Descriptors *Changes in appearance over time are also relevant for action recognition.*
Yilmaz and Shah [2005], Peursum et al. [2004], Peursum et al. [2005], Jiang et al. [2006], Kosta et al. [2006], Filipovych and Ribeiro [2008], Wang et al. [2009], Dhillon et al. [2009]	Datong et al. [2008], Abdelkader et al. [2008], Cuntoor et al. [2008], Hu et al. [2008a], Hu et al. [2008b]	Bobick and Davis [2001], Efros et al. [2003], Hatun and Duygulu [2008], Hu et al. [2009], Wong et al. [2007], Thurau and Hlavac [2008]	Blank et al. [2005], Mokhber et al. [2008], Cuntoor [2006], Fathi and Mori [2008], Ke et al. [2005], Ke et al. [2007a], Laptev and Perez [2007]

3.3 Approaches Based on Models of the Moving Objects

In this section, approaches which do assume that specific objects are in the scene under constrained conditions are discussed according to the structure proposed in upper part of Figure 3.2. This category of approaches relies on models for those objects assumed to be performing the actions, models that can be either explicit (parametric) or implicit.

As indicated in Figure 3.2, approaches based on object models are those that appeared first in the literature, out of traditional motion capture research. Some recent and/or classical papers of each branch are cited as needed throughout the following text, but the list presented is not meant to be exhaustive, since those approaches are broadly covered in previous surveys (Section 3.1). Therefore, this section is mainly aimed at providing a high-level overview of that category of approaches, detailing it enough to give the reader some perspective on the evolution of the area towards less constrained techniques.

3.3.1 Using Parametric Object Models

The approaches based on parametric object models are the ones more directly related to motion capture techniques. In such approaches, moving objects (e.g., the human body, the hand, cars in a parking area) are assumed to follow a specific model and visual data is matched against that model to infer its parameters. In the action modeling step, different sets of values for the model parameters are then associated with different actions.

Approaches based on parametric models can be split into two major subgroups: those which parameterize the object internal structure, and those which ignore such internal structure and instead, parameterize only objects trajectories.

Models of the Objects Internal Structure

These approaches, which are among the earliest techniques developed for action recognition, are mostly derived from motion capture techniques. Thus, they start by defining a detailed model of the internal structure of a pre-defined moving object and then adjusting the visual data to that model. The most commonly modeled objects are the entire human body as well as body-parts, such as hand models aimed at gesture recognition, for example. A classical example of human-body model (the so-called stick model) can be seen in Figure 3.2(a), from Yilmaz and Shah [2005].

Strictly speaking, a detailed explicit model would require a great amount of 3D data, like in Yilmaz and Shah [2005], thus leading to a high computational complexity. In addition, in most real-world scenarios, 3D data is simply not available. Hence, a number of approaches avoid that requirement by using simplified models. In Peursum et al. [2004, 2005], for example, a simplified stick model is obtained from silhouettes. Poses are modeled based on a few points from the body in Jiang et al. [2006]. In Kosta et al. [2006] the human body is considered as a cooperative team of agents where each team member is a limb of the body. In Filipovych and Ribeiro [2008] and Wang et al. [2009] constellation models are employed to describe human poses. In Dhillon et al. [2009], body parts are tracked using mixture particle filters and clustering the particles locally.

Trajectory Models

In trajectory-based approaches, the global motion of the objects is considered the only relevant information for action recognition. In other words, the internal state of the moving objects is ignored and such objects are represented mainly by their position tracked over time. Actions are then modeled by trajectories parameters, which, in turn, can come from a number of different trajectory models. Trajectory-based approaches are very common in surveillance scenarios – such as the one depicted in Figure 3.2(b), from Datong et al. [2008]. A number of surveys specifically devoted to them were already discussed in Section 3.1.

In Abdelkader et al. [2008], for example, it is argued that activities can be modeled by any representative shape associated with the activity to be modeled, and as an example case, the shape of the trajectories of a set of points associated with the moving object is analyzed. Instead, a large number of proposals analyze the trajectory of a unique point, as in Cuntoor et al. [2008], which is aimed at identifying common office activities by the analysis of hand trajectories. In Hu et al. [2008a] and Hu et al. [2008b], the focus is on detection of abnormal events in crowded scenes. In such scenes, tracking the objects of interest is particularly challenging. To overcome this, global motion fields are analyzed in order to discover *super-tracks*, which are intended to capture predominant motion patterns that are then used to model events.

3.3.2 Using Implicit Object Models

In this class of approaches, the area around the moving object — like a silhouette or a bounding box, for example — is detected and submitted to some kind of global description. This line of work assumes that explicit details of the object structures are not

necessary for action recognition. Rather, the global features of a ROI defined around the object implicitly capture its model, at a lower cost. The lower computational effort offered by this basic idea gave rise to a variety of similar approaches, which can be distinguished between those using 2D templates and those exploiting spatio-temporal 3D templates.

Implicit Object Models Based on 2D Templates

A landmark paper using implicit object modeling as the basic representation is Bobick and Davis [2001]. In this paper, two different 2D templates computed from extracted silhouettes is proposed: (a) Motion Energy Image (MEI), which is a binary image indicating where the motion occurred during the sequence; (b) Motion History Image (MHI), which is a gray level image where brighter pixels indicate the recency of motion (in other words, the brighter the pixel, the more recent the motion occurred there). Both MEI and MHI images are described by seven Hu moments, which are meant to carry a coarse shape description which is invariant to scale and translation (Figure 3.2(c)). The MEI/MHI representation proposed by Bobick and Davis [2001] became the basis of a great number of extensions and variations, mostly applied to scenarios with relatively stable backgrounds (like the work of Hu et al. [2009] aimed at surveillance applications).

Another classical approach using 2D templates as primary video representations appears in Efros et al. [2003], which addresses the problem of recognizing human actions from medium resolution videos. This approach relies on the detection and stabilization of a bounding box containing the human figure. The description of such boxes is based on the optical flow projected into motion channels, which are blurred with a gaussian filter to reduce the sensitivity to noise which is typical of optical flow estimations. Such motion descriptors are later used by several other authors (see Wang et al. [2007] and Fathi and Mori [2008], for instance).

In Wong et al. [2007], the internal part of previously detected motion regions are described by BoVF, a statistical representation based on interest points that is further detailed in Section 3.4.2.

Along with the global spatial descriptors, the information encoded in the sequential nature of video is explicitly taken into account by some authors relying on 2D templates. In Hatun and Duygulu [2008], a bounding box centered at the moving body is described based on radial HoG. Such histograms are then clustered to create a codebook of poses, based on which each video is described by two alternative representations, namely: bag-of-poses and sequence-of-poses. A similar approach – using

Non-negative Matrix Factorization (NMF) to build a bag-of-poses representation – is presented in Thurau and Hlavac [2008].

Implicit Object Models Based on Spatio-temporal 3D Templates

In this category, actions are represented as 3D volumes in space-time. Such spatio-temporal volumes are created by aligning and stacking 2D information (e.g., silhouettes, contours, bounding boxes). The exploration of space-time volumes built on silhouettes for action recognition was first proposed in Blank et al. [2005]. In their proposal, the properties of the Poisson equation are used to create a representation in which the values reflect the relative position of each internal position in the volume (Figure 3.2(d)).

The principle behind such approaches is that spatio-temporal volumes contain both static and dynamic information and are thus better suitable as representations for action recognition. Similarly to what happened with 2D template-based approaches, the initial idea of Blank et al. [2005] is further explored in a number of subsequent papers, which describe the spatio-temporal volume using different techniques. For instance, Mokhber et al. [2008], characterize the spatio-temporal volumes by their 3D geometric moments. In Cuntoor [2006], it is proposed an algebraic technique for characterizing the topology of those volumes. Finally, the representation used by Fathi and Mori [2008] can be considered an extension of the work of Efros et al. [2003] to a 3D spatio-temporal volume.

In Ke et al. [2005] and Ke et al. [2007a], the authors explore space-time volumes at a smaller scale. Videos are over-segmented in space-time, creating micro-volumes, which are described based on optical flow information. They are then compared against manually segmented action templates, by a shape-matching technique adjusted to deal with over-segmentation.

In order to distinguish between drinking and smoking actions, Laptev and Perez [2007] use densely sampled HoG and HoF 3D descriptors computed over manually cropped regions around people's faces, used as input to a cascade of weak classifiers learned by AdaBoost.

3.4 Approaches Based on Global Statistics

Approaches which rely on the detection of moving objects share the drawback of depending on computer vision tasks – such as background segmentation and tracking – which are themselves open research issues. The lack of general solutions to those

tasks leads to an excessive number of assumptions about what is in the scene, which ultimately makes such approaches applicable only to very constrained scenarios.

To cope with more realistic and unconstrained settings, different approaches make no assumption on the presence of any specific object in the scene, thus making object detection unnecessary. Instead, those approaches compute global statistics on different data. Statistics on *low-level features* (such as color, texture and optical flow, for example) can be computed either as generic descriptors or guided by specific information about the application domain, as further discussed in Section 3.4.1. Mid-level *Local descriptors* built on low-level data around selected points gave rise to an important branch of approaches based on Bag-of-Visual-Features (BoVF), which are essentially histograms of quantized local descriptors. BoVF-based approaches are thoroughly discussed in Section 3.4.2. Although BoVF-based approaches dominated the scenario of statistical approaches in recent years, alternative proposals to gather information coming from local descriptors can be found and are discussed in Section 3.4.2. A third research line exploits the probabilities of high-level semantic concepts (e.g., sky, airplane, people) appearing in a video to infer the action taking place in it. These approaches are detailed in Section 3.4.3

3.4.1 Using Statistics of Low-level Features

In this category of action recognition approaches, low-level features of the video are statistically summarized and such summary is used as the video representation. Since the direct usage of low-level features is prone to suffer more intensely the effects of the semantic gap, some authors use previous knowledge about the application domain to guide the choice of features. Generic features without specific links to the application domain have also been exploited, although such approaches tend to be focused on a handful of constrained settings.

Low-level Statistics Guided by Domain Knowledge

A combination of low-level features and some previous domain knowledge is common in scenarios where the possible backgrounds are limited in number and have distinct global appearance. In professional sport videos, for instance, camera effects are commonly related to specific events. This fact is used in Xie et al. [2004], in which dominant color ratio and motion intensity are computed to segment soccer videos between play and break intervals.

Another application for approaches based on low-level statistics is explored in Papadopoulos et al. [2008], which use global motion information to identify generic,

coarse-grained events in news videos, like *anchor*, *reporting*, *reportage* and *graphics*.

Many approaches based on low-level statistics appear as part of multimodal frameworks (see, e.g., Tien et al. [2008]), in which audio-visual features are mixed with high-level information to detect events in tennis games. The full exploration of multimodal frameworks is outside the scope of this paper, although we point the reader to Snoek and Worring [2005] for a related survey.

Generic Low-level Statistics

A variety of approaches for computing generic global statistics based on low-level features, which do not rely on domain information, have been proposed in the recent literature. In most cases, they relate to constrained applications.

In Lavee et al. [2007], a framework aimed at searching for suspicious actions represents candidate video segments by histograms of intensity gradients both in spatial and temporal dimensions, over four different temporal scales. In Ermis et al. [2008], the typical dynamics of a surveilled environment is captured by statistical analysis of a 2D field containing the maximum activity of pixels. From that field, a model for normal behavior is produced, allowing comparisons with other videos so as to detect abnormal activities. Surveillance scenarios are also the focus of Ma and Cisar [2009], which propose a dynamic texture descriptor based on local binary patterns extracted from the three orthogonal planes formed by the spatial and temporal axes. The videos are represented by sequences of such descriptors computed over subsequent spatio-temporal subvolumes.

In Zelnik-Manor and Irani [2001] and Zelnik-Manor [2006], a generic approach for what is called Smart Fast-Forward (SFF) is presented. Their approach is based on the absolute values of normalized gradients computed over all space-time points, extracted in a temporal pyramid, to cope with different temporal scales. Points with gradients below a threshold are ignored in order to save time, and the remaining ones are described by the gradient components in x, y and t directions, for all temporal scales considered.

Similarly, in Shechtman and Irani [2005], underlying motion patterns are applied to identify video segments similar to a query sequence. This is done by computing the correlation of such motion patterns in the query video segment with a larger video sequence. The peaks in the correlation surface correspond to similar sequences. In this approach, the motion is estimated from the gradients inside small spatio-temporal patches or cuboids, instead of relying on expensive flow computations.

The proposal of Shechtman and Irani [2007] computes the similarity between

images or videos by matching local self-similarities. Those are computed at pixel level, taking into account the similarity between a small patch around the considered pixel and a larger region surrounding it.

Also aimed at SFF applications, Seo and Milanfar [2009] propose to recognize actions from a unique example, by using local regression kernels based on weights computed on the video pixels and their neighbors both in space and time.

In Li [2007], the optical flow computed for the entire video is represented by magnitude and orientation. A histogram is built on the quantized orientation, using only pixels for which the flow magnitude is above a certain threshold. Also, the flow of the considered pixels is weighted by their magnitude. The normalized histograms for the training sequences for each class are submitted to PCA for dimensionality reduction.

In Ahammad et al. [2007], motion vectors from the compressed-domain are used to estimate motion fields, which are then submitted to a hierarchical agglomerative clustering algorithm, in order to create an organizing hierarchy for videos which are presumed to be based on actions.

In contrast to the BoVF-based approaches (see Section 3.4.2 below), Wang et al. [2007] argue that human actions should be characterized by large-scale features instead of local patches. Therefore, the authors consider the frame as the basic unit for initial description, which is made in terms of the motion descriptors proposed by Efros in Efros et al. [2003]. Their "visual vocabulary" is then built on those global frame features, whose space is quantized by the k-medoid clustering algorithm. Finally, each video sequence is represented in terms of the frequency of such "frame-words".

In Ning et al. [2009], a hierarchical space-time model is implemented in two layers: the bottom layer of features composed of a bank of 3D Gabor filters; the second layer in the proposed hierarchy are histograms of Gabor orientations. This proposal is based on that of Serre et al. [2007] for object recognition, which tries to mimic organic visual systems, which are seen as being composed of two kinds of brain cells with different roles in the recognition process.

Finally, in Chaudhry et al. [2009], the temporal evolution of Histograms of optical Flow (HoF) features gathered from each frame are modeled by a non-linear dynamical system.

3.4.2 Approaches Based on Statistics of Local Descriptors

Last section discussed the first attempts at avoiding the constraints imposed by object models for action recognition. However, most of those approaches either rely on domain

knowledge or are focused on constrained settings or databases, like surveillance or Smart Fast-Forward (SFF) applications. Another drawback of those approaches is that, being based on dense low-level features, they demand great computational effort.

To mitigate those drawbacks, approaches based on mid-level local descriptors, mostly computed on a (potentially small) number of interest points emerged as a promising trend for action recognition. More specifically, approaches based on histograms of quantized local descriptors – known as Bag-of-Visual-Features (BoVF)[1] – have shown to provide consistently good results reported by a number of independent authors in a variety of scenarios, including datasets composed of professional and amateur realistic videos.

Despite the success of BoVF-based approaches, there are a few other strategies to gather information from local descriptors. These alternative strategies are gathered in a separate category in the proposed framework.

Using Bag-of-Visual-Features (BoVF)

BoVF representations are inspired by traditional textual Information Retrieval (IR) techniques, in which the feature vectors that represent each text document in a collection are histograms of word occurrences Baeza-Yates and Ribeiro-Neto [1999][2]. Such representation is referred to as Bag-of-Words (BoW), in order to emphasize that it is comprised of orderless features.

A remarkable difference in the analogy between BoVFs and BoWs is the need to define what constitutes a *visual word*. Such "definition" is achieved in practice by a process called *vocabulary (*or *codebook) learning*, consisting of the quantization of the descriptors' feature space, typically computed by clustering. A detailed introduction on how BoVF representations are build both for images and videos can be found in Lopes et al. [2009c].

BoVF-based approaches have been first applied to object classification and have proved very robust to background clutter, occlusion and scale changes, indicating their potential for challenging object recognition settings (Agarwal and Awan [2004], Lazebnik et al. [2006], Zhang et al. [2007b], Jiang et al. [2007], Wong and Cipolla [2007], Sun et al. [2009a]). BoVF and its variations have demonstrated similar strengths when applied to action recognition, thus becoming, by far, the most common base representation found on recent proposals.

[1] Due to the lack of standard terminology, those approaches have also been denominated bag of visual words, bag of keypoints, bag of features or bag of words in the literature.

[2] In fact, each histogram bin reflects not a single word, but a family of words represented by their roots.

The relevance of BoVF-based action recognition is reinforced by the fact that those schemes became a common testbed for several spatio-temporal points detection and description algorithms. The work of Schindler et al. [2008], for example, compares different alternatives for interest point detectors/descriptors applied in a classic BoVF representation for Internet videos. Similar comparisons can also be found in Wong and Cipolla [2007], Klaser et al. [2008], Willems et al. [2008], Kaiser and Heidemann [2008] and Rapantzikos et al. [2009].

To the best of our knowledge, the approach proposed by Schuldt et al. [2004] is the seminal work on BoVF techniques applied to action recognition. For the low-level features, the spatio-temporal interest points proposed in Laptev and Lindeberg [2003] are described by spatio-temporal jets. K-means clustering is applied to create the quantized vocabulary, based on which the histogram of local features is computed. This is also the work which introduced the Royal Institute of Technology – in Swedish (KTH) action database, which later became a *de facto* standard for action recognition algorithms. The work described in Schuldt et al. [2004] is extended in Laptev et al. [2007], which proposes a mechanism for local velocity adaptation aimed at compensation of camera motion that could affect local measurements.

The work of Dollar et al. [2005] extends previous work on object recognition based on sparse sets of feature points. The interest points selection method applied in this work is based on separable linear filters. Three descriptors for the cuboids delimited around the selected points are tested: *normalized pixel values*, *brightness gradients* and a *windowed optical flow*. PCA is used for dimensionality reduction of the point descriptors and a typical BoVF signature is then built on them. The k-means clustering algorithm is used for defining the dictionary.

Another BoVF approach is proposed by Scovanner et al. [2007], this time based on an extension of the SIFT descriptor Lowe [1999]. The new descriptor adds temporal information, extending the original SIFT descriptor to a 3D space. Instead of using the SIFT detector, points are selected at random. Histograms built on a codebook created with k-means are the initial signatures. Then, to create an enhanced representation, a criteria based on the co-occurrence of visual words is applied to reduce the vector dimension. In other words, those visual words which co-occur above a certain threshold are joined.

In Ning et al. [2007] the local descriptors are based on the responses to a bank of 3D Gabor Filters, followed by a MAX-like operation. Such features are computed on patches delimited by a sliding window and described by histograms generated by the quantization of the orientations in nine directions. The quantization of those histograms into a codebook is learned from a gaussian mixture model.

In Niebles et al. [2008], a BoVF representation based on the features proposed by Dollar Dollar et al. [2005] is used together with generative models – unlike previously discussed methods which are based on discriminative ones – for action recognition. Two methods borrowed from traditional textual Information Retrieval research are examined: probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA).

In Marszalek et al. [2009], a joint framework using BoVFs both for scene and actions recognition is proposed. The underlying assumption is that scenes can provide contextual information to improve the recognition of some actions classes. Initially, movie scripts provide for automatic annotation of scenes and actions. Text mining is then applied to discover co-occurrences between scenes and actions. Finally, separate SVM models based on BoVFs are learned using the same approach described in Laptev et al. [2008].

BoVF Variations

A number of variations over the typical BoVF scheme have been proposed, mostly aimed at dealing with specific recognized limitations of pure BoVF-based approaches, the main ones being the lack of structural information and the poor quality of the visual vocabulary.

In Zhao and Elgammal [2008], the lack of structural information of BoVF representations is addressed in a rather direct manner: each frame is subdivided into cells, over which BoVF based on Dollar's features Dollar et al. [2005] are computed. Additionally, motion features from neighbor frames are used in a weighted scheme which takes into account the distance of the neighbor from the actual frame. A spatial-pyramid matching, similar to the one used in Lazebnik et al. [2006], is then applied to compute the similarity between frames. Finally, frames are classified individually, and a voting scheme decides the final video classification.

In Laptev et al. [2008], it is proposed a BoVF representation built from the STIP presented in Laptev and Lindeberg [2003], but without scale selection. The points are described by Histograms of Gradients (HoG) and Histograms of optical Flow (HoF), computed over the spatio-temporal volumes positioned around them. To add some structural information, each video volume is subdivided by a grid, so that at recognition time different configurations for the grids are considered, using a multi-channel SVM classifier.

Most authors working on BoVF-based approaches for actions recognition learn the vocabulary by using the k-means clustering preceded by a PCA-based dimensionality

reduction. Nevertheless, some alternatives to vocabulary learning have been proposed, either by enhancing the vocabulary delivered by k-means or by applying alternative clustering techniques.

The work of Liu and Shah [2008] merges k-means results to produce an enhanced vocabulary, indicating that better vocabularies can have a significant impact on recognition. Using features similar to those of Dollar et al. [2005], they propose a Maximization of Mutual Information (MMI) criteria to merge the cuboid clusters output by k-means. Those new clusters are called Video Words Clusters (VWC). Additionally, two approaches to add structural information are explored: spatial correlogram, with the distances quantized in a few levels and the Spatio-Temporal Pyramid Matching (STPM) of Lazebnik et al. [2006].

Another example of the impact of a better vocabulary is found in Jiang and Ngo [2009], where the authors proposed an enhanced BoVF in which the relationships among visual words are explored. This is pursued with the aid of a visual ontology inspired on WordNet Fellbaum [1998], a textual ontology extensively applied for text retrieval. Their visual ontology is built by applying agglomerative clustering to the visual words previously discovered by k-means. From that, it is possible to compute the specificity (i.e., the depth in the ontology tree), path length (i.e., the number of links in the path between two words) and information content (relative to the probability of word occurrences). Those precomputed values are used in a soft-weighting scheme aiming at better evaluating the significance of each word.

In Liu et al. [2009], a BoVF combining dynamic and static local features is proposed to address action recognition in YouTube videos. Static information captured by three different interest point detectors are described by SIFT descriptors. Motion is collected by Dollar's interest points Dollar et al. [2005], described by gradient vectors. The spatio-temporal distribution of motion features is used to localize coarse ROIs, which are used together with the PageRank algorithm, for prunning spurious features. In addition to such motion-guided feature selection, the authors propose a procedure to create a semantic visual vocabulary, which involves enhancing the result of k-means by using a technique based on the KL-divergence algorithm.

In Datong et al. [2008], a modified version of k-means which takes into account the spatial localization of the interest points is applied to form the codebook. Additionally, a 3D extension of the Harris detector alternative to that of Laptev Laptev and Lindeberg [2003] is proposed, aimed at selecting a denser sampling. Cuboids defined around the detected points are described in terms of shape and motion.

Finally, in Zhou et al. [2008], the visual word distribution is described by Gaussian Mixture Models (GMM) of SIFT descriptors. These GMMs are specialized for each

video clip, and a kernel for video comparison is built on the Kullback-Leibler divergence (Moreno et al. [2003] *apud* Zhou et al. [2008]).

Alternative Strategies to Capture Information from Local Descriptors

Despite the great success of BoVF-based approaches and their variations, achieving high recognition rates in truly realistic databases remains an open challenge. A number of authors have been trying alternative ways to collect relevant information out of local descriptors.

In Oikonomopoulos et al. [2005], the salient-points detected at peaks of activity are clustered into salient regions, whose scales are correlated to the motion magnitude. Noisy interest points are discarded, so the videos are described by remaining points inside detected salient regions. Since such representations do not have the same dimension for all video sequences, the Chamfer distance is used as the kernel for a Relevance Vector Machines (RVM) classification algorithm. A space-time warping adjustment scheme is applied to deal with varied execution speeds. Some variations to this overall scheme are presented in Oikonomopoulos et al. [2006].

In Savarese et al. [2006], it is argued that correlograms can capture the spatial arrangement of codewords in the case of object classification. This is applied in Savarese et al. [2008], in which an extension of the spatial correlatons (quantized in *correlograms*) is proposed for action recognition. Rather than building a histogram of interest-points, as in typical BoVF approaches, action is modeled as a collection of space-time interest points where each interest point has a label from the vocabulary of video-words. So, the video sequences are composed of sets of video words and their spatio-temporal relations are described in form of spatio-temporal correlatons. Actions are modeled by estimating – with pLSA – the codewords distribution for each particular class.

In Nowozin et al. [2007], it is observed that the lack of structural information also means the absence of temporal sequencing. The representation proposed uses Dollar's features Dollar et al. [2005], but the histogram built on them is made up of temporal bins. PrefixSpan algorithm is used for mining frequent sequences and the LPBoost algorithm is used to identify the most discriminative ones.

In Gilbert et al. [2008], dense corner features are hierarchically grouped in space and time to produce a compound feature set. Data mining is applied to group features in multiple stages, from the initial low-level features until a higher level in which the relative positions of groupings are used. As in Lopes et al. [2009c], 2D features are collected from the planes defined by the coordinates (x, y) – the frames – (x, t) and (x, t), which are, in the former case, considered as distinct channels. Motion is

captured by dominant orientation and points are described only by their scale, corresponding channel and orientation, instead of more complex descriptors like SIFT or STIP. Transaction vectors are built based on neighbor interest points and the Apriori algorithm is applied to the transactions in order to find association rules between transactions vectors and actions.

In Ryoo and Aggarwal [2009], a matching kernel function for comparing spatio-temporal relationships among interest points is proposed for detection and localization of multiple actions and interactions in unsegmented videos. First, interest points are detected and described as usual. Afterwards, pairwise relationship predicates are used to describe the structural relations. Temporal relations are described by Allen's taxonomy (*equals*, *before*, *meets*, *overlaps*, *during*, *starts* and *finishes*) Allen and Ferguson [1994] *apud* Ryoo and Aggarwal [2009], with respect to the interval limits given by the volume patch dimensions projected onto the temporal axis. Similar spatial predicates are created, so that temporal and spatial 3D relationship histograms aimed at capturing both appearance and point relationships in the video can be computed. Finally, the proposed matching kernel captures the similarity between two histograms by their intersection.

The work presented in Uemura et al. [2008] proposes a video representation in the form of a vocabulary-tree, based on the outputs of several interest point detectors plus a dense sampling for action recognition and localization. Motion compensation is achieved by using previously tracked features to perform the segmentation of the image into motion planes. Such a segmentation is performed by an initial color-based segmentation followed by homography computation using RANSAC. The homography is then used to correct the motion of the features inside each dominant plane.

The fact that humans can recognize actions just by observing some tracked points has been explored in several approaches relying on trajectory models of points placed at specific body parts. The work of Sun et al. [2009a] extends this notion into a more generalizable approach, in which the authors propose to gather information about the spatio-temporal context of tracked SIFT points in a hierarchical three-level scheme. In the first level – the point-level context – local statistics of gradients along point trajectories are computed. In the second level – the intra-trajectory context – the dynamic aspects of those trajectories in the spatio-temporal space are considered. Finally, in the coarser level – the inter-trajectory context – the information about the spatio-temporal co-ocurrences of trajectories distributions is collected.

In a similar vein, the tracks of a set of features – detected and tracked by the algorithm proposed in Lucas and Kanade [1981], but with weaker constraints – is employed by Messing et al. [2009]. Such trajectories are described by the history of

their quantized velocities.

3.4.3 Using Concept Probabilities

Using similar ideas from CBVR systems Snoek and Worring [2008], some authors have proposed to use higher level concepts as the building blocks of video representations aimed at action recognition.

In Ebadollahi et al. [2006b], 39 semantic concept detectors from Large-Scale Concept Ontology for Multimedia (LSCOM)-lite – SVM classifiers based on raw color and texture features Naphade et al. [2005] – are applied to video I-frames. Then, the trajectories of those concepts in the concept space are analyzed by Hidden-Markov Models (HMM), one for each concept axis. Their work reinforces the results of Kennedy [2006], providing additional evidence that, for some concepts, the dynamic information is essential. The authors found that dynamic information enhanced recognition results of the following concepts: *riot*, *exiting car* and *helicopter hovering*.

In Haubold and Naphade [2007], Moving Picture Experts Group (MPEG) motion vectors are summarized in a motion image which describes the global motion pattern of a video shot. Motion images are combined with color and texture features and used as input for several weak SVM classifiers. The output of such classifiers are fused together to compose the video feature vector. The approach is tested on the TREC Video Retrieval Evaluation (TRECVID)-2005 dataset, for selected dynamic concepts only, comparing favorably with results based on motion direction histograms and motion magnitude histograms.

In Xu and Chang [2008], 374 concepts are selected from the LSCOM ontology to be detected by three different SVM detectors based on histograms of low-level features (grid color moments, Gabor textures and edge direction histograms). The results of those classifiers are fused together in order to produce scores for each concept. Variations in the duration of action clips are dealt with by applying the Earth's Mover Distance (EMD) in multiple temporal scales.

A framework for event detection presented in Wang et al. [2008] starts by the application of BoVF-based approach to detect a number of concepts. Relative motion of keypoints between successive frames is used to aid the spatial clustering of visual words. A visual word ontology is then built based on the output of the spatial clustering, in order to take into account the correlation of visual words potentially related to the same object or object parts. The final representation is a collection of BoVFs built on those roughly segmented regions.

3.5 Hybrid Approaches

This section is devoted to action recognition approaches that fuse information coming from both object models and global statistical representations. From Figure 3.2, it is possible to notice that hybrid approaches come up almost concomitantly with those approaches based on global statistics, in an attempt to draw on the advantages of both kinds of representations. It is worth noticing though, that from the point of view of generalization ability, such mixed approaches are limited by the representation whose computation imposes stronger constraints.

In Niebles and Li [2007], it is proposed a hybrid approach where constellation models are used to add geometric information to the classical BoVF representation. This is done by modeling actions within a two-layered hierarchical model. The higher layer is comprised of selected body parts, which are then described as BoVFs. The BoVF-based system proposed is based on Dollar's features Dollar et al. [2005], together with sampled edge points described by shape context.

The work presented in Mikolajczyk and Uemura [2008] mixes low-level, local descriptors and shape-based representations. They build several vocabulary trees on points selected by five different 2D point selectors. To include dynamic information, motion maps are obtained from a Lucas-Kanade optical flow computation, and motion is represented by velocity maps between pairs of frames. A technique for compensation of camera motion based on a global similarity transformation is also presented and applied. A star-shape model – aimed at coarsely capturing some structural information of the moving object – is used to guide the process.

In Junejo et al. [2008], the concept of Self-Similarity Matrix (SSM) is introduced to build video representations aimed at action recognition. An SSM is a table of distances between all video frames. Although this definition can be applied for any feature type, in Junejo et al. [2008], they are computed over trajectories of human joints, which are fused together with those computed from HoG and HoF features. The final descriptor is obtained by considering the SSM sequences as images and splitting them into patches, which are described by histograms of gradient directions.

The proposal of Bregonzio et al. [2009] works on clouds of interest points collected over different time scales. The distribution of such clouds in both space and time is described by global features. To compose the clouds, they propose a spatio-temporal interest point detector that collects dense samplings of interest points. In order to avoid too many spurious point detections, the moving object is coarsely separated from the background.

In Sun et al. [2009b], BoVFs of 2D and 3D SIFT feature descriptors, extracted on

2D SIFT interest points, as well as Zernike moments, are applied to both frames and MEI images. The extraction of those features are guided by frame subtraction, which provides a coarse motion-based segmentation. Feature fusion is achieved by simple concatenation of the resulting four descriptors.

The work of Oikonomopoulos et al. [2009] proposes another hybrid approach for both recognition and detection of actions in unsegmented videos. Visual codebooks are class-specific and take co-occurrences of visual words pairs into account. The positions of these visual words in relation to the object center are used to model the actions, which therefore implies the need for object segmentation. In addition, spatio-temporal scale adjustments are done manually for training. Finally, a framework for voting over time, which is based on optical flow and appearance descriptors, is proposed for action segmentation.

3.6 Concluding Remarks

This survey attempts to summarize the efforts of the academic community at the task of recognizing human actions from monocular videos, with emphasis on recent approaches. It proposes a new organizing framework, based on the representations chosen, and therefore, on their underlying assumptions. This organization allows to categorize the newest approaches smoothly alongside the more traditional ones. It also allows to compare and contrast different methods based on their constraints, which, we hope, enables a principled selection of a method, given the application domain. We observe that there is a correlation between our classification criteria and the chronology of methods, indicating a trend toward progressively weakening the constraints imposed on video content. Nevertheless, generalization comes at a cost, and hybrid methods begin to attempt drawing on the advantages of both identified major lines.

Figure 3.2 together with Tables 3.1, 3.2 and 3.3 summarize the survey. As already discussed throughout the text, approaches relying on object models to describe video content have the drawback of imposing a number of constraints on the action scenario. Such constraints are rarely met in feature movies or user-generated videos found in video sharing systems (e.g., YouTube), thus pushing the research in action recognition towards more general approaches. It is important to notice, though, that some approaches based on object models have proved successful in realistic but restrict application domains, like surveillance and Human-Computer Interaction (HCI), for example. In fact, provided that their constraints can be guaranteed, those approaches should be considered as potential choices in those cases where real-time processing is

a requirement. In particular, approaches based on implicit models built on 2D descriptors – for instance, like those based on MEIs and MHIs Bobick and Davis [2001] – tend to be quite efficient. In addition, the advances in the state-of-the-art of preprocessing techniques like segmentation and tracking can turn some approaches based on object models better suited for realistic environments, possibly giving rise to new hybrid approaches.

The greater focus of the survey on BoVF-based approaches emerges naturally from their potential on the field, making it a promising direction to pursue in the search for effective solutions for recognizing human actions in scenarios of realistic videos. It also reflects the focus selected for this work, which is based on BoVF representations.

Table 3.2. Summary of Action Recognition Approaches Based on Global Statistical Representations (table best viewed in color).

Basic Representation – Statistics *Main Assumption: Global statistics capture relevant information for action recognition.* [Related Papers by Subcategory (below)]				
Low-level Features *Low-level features can indicate actions being performed.*		Local Descriptors *Mid-level local descriptors are better suited to capture relevant information.*		Concepts *Concepts occurrences can indicate the actions being performed.*
Domain-Oriented *Domain information must guide the choice of relevant low-level features.*	Generic *Generic low-level features are able to capture relevant information.*	BoVF *A histogram of quantized local descriptors can be associated with actions.*	Other *Other ways than BoVF can better capture relevant information from local descriptors.*	
Xie et al. [2004], Papadopoulos et al. [2008], Tien et al. [2008], Snoek and Worring [2005]	Lavee et al. [2007], Ermis et al. [2008], Ma and Cisar [2009], Zelnik-Manor and Irani [2001], Zelnik-Manor [2006], Shechtman and Irani [2005], Shechtman and Irani [2007], Seo and Milanfar [2009], Li [2007], Ahammad et al. [2007], Wang et al. [2007], Ning et al. [2009], Chaudhry et al. [2009]	Lopes et al. [2011], Lopes et al. [2009c], Schuldt et al. [2004], Laptev et al. [2007], Dollar et al. [2005], Scovanner et al. [2007], Ning et al. [2007], Niebles et al. [2008], Marszalek et al. [2009], Zhao and Elgammal [2008], Laptev et al. [2008], Liu and Shah [2008], Datong et al. [2008], Zhou et al. [2008]	Oikonomopoulos et al. [2005], Oikonomopoulos et al. [2006], Savarese et al. [2006], Nowozin et al. [2007], Gilbert et al. [2008], Ryoo and Aggarwal [2009], Uemura et al. [2008], Sun et al. [2009a], Messing et al. [2009]	Ebadollahi et al. [2006b], Haubold and Naphade [2007], Xu and Chang [2008], Wang et al. [2008]

Table 3.3. Summary of Hybrid Action Recognition Approaches (table best viewed in color).

Basic Representation – Hybrid *Main Assumption: combinations of different approaches produce enhanced recognizers.* [Related Papers (below)]
Niebles and Li [2007], Mikolajczyk and Uemura [2008], Junejo et al. [2008], Bregonzio et al. [2009], Sun et al. [2009b], Oikonomopoulos et al. [2009]

Part II

Contributions

Chapter 4

Proposed Solution

In this work, we propose the implementation and experimental validation of a new methodology to collect and fuse contextual information with BoVF representations for videos. As it will be detailed in Section 5, the proposed methodology provides two main original contributions: firstly, it describes context in a higher semantic level than previous approaches, by the application of state-of-the-art semantic video retrieval techniques. Such goal requires the training of a number of concept detectors.

That requirement, by its turn, touches on an important drawback of supervised ML techniques underpinning most of the work in image and video understanding: the lack of enough labeled training samples. The solution proposed to overcome this drawback leads to our second major contribution, which is the usage of TL theory in order to benefit from existing clean annotations on auxiliary external databases, independent from the target action database.

4.1 Justification

4.1.1 Contextual Information in Action Recognition

In Marszalek et al. [2009], the usage of contextual information in aiding to recognize related actions is addressed. In that work, relevant scene types are learned from the texts of movies scripts. Scene and human-actions are then described as BoVFs and SVM classifiers are trained separately for both kinds of concepts. Finally, the correlation between scenes and actions is explored to enhance the recognition results both for scenes and actions. In their case, the contextual information is provided by a global BoVF-based description of the entire scene, instead of particular objects or concepts.

In Ebadollahi et al. [2006b], the context for events selected from the LSCOM

annotations is captured by the trajectories of videos in the concept space, which is defined by a number of concepts selected from LSCOM-lite annotations (Kennedy and Hauptmann [2006]). Concept detectors are trained on low level features like color, texture, motion and edges, and the trajectories themselves are analyzed by HMM, whose scores are used to form a feature vector for the final SVM classifier. As it will be seen in Chapter 3, low-level features are not the best ones for action classification, leading to poorer results when compared to BoVFs and other representations based on interest points. Moreover, in Ebadollahi et al. [2006b], no direct dynamic information is taken into account.

The work of Sun et al. [2009a] explores context at lower levels. In that work, contextual information is encoded by different contextual descriptors computed on trajectories of SIFT points. Such descriptors are combined with different spatio-temporal grids in a multi-channel kernel scheme similar to that applied in Laptev et al. [2008].

Interestingly enough, the comparison among several 3D point detectors and descriptors performed by Wang et al. [2009] concluded that, except for the KTH database, regularly spaced dense samplings of points perform better at action recognition than interest points. This result reproduces similar ones obtained for scene recognition by Fei-Fei and Perona [2005], and can be considered as additional – though more indirect – indications of the importance of context in realistic videos, since denser samplings, by covering larger areas, should be better able to capture contextual information. The distinct result for KTH can be explained by the fact that it is a controlled database, whose backgrounds are almost completely uniform. Hence, there is virtually no context to be described in KTH.

4.1.2 Dealing with Scarce Training Data

To manually collect concept examples from realistic videos is a time-consuming and error-prone task. This is a serious bottleneck to research related to video understanding, since the large intra-class variations of such videos demand training sets large enough to properly encompass those variations.

In the pursuit of finding a scalable way for collecting training samples for different actions in movies, both Laptev et al. [2008] and Marszalek et al. [2009] use textual information coming from movie scripts to automatically segment actions and scenes samples. Their procedure is further refined in Duchenne et al. [2009], in which a clustering algorithm similar to k-means is applied to temporal sub-windows of the initial video clip, in order to narrow it into a more precise action location in time.

In Ulges et al. [2009], the issue of collecting a large-enough amount of training

data for high-level video retrieval is addressed by the usage of videos collected from Youtube[1] filtered by pre-defined categories and tags.

The resulting annotations achieved in those cases, though, are quite noisy. Both the experiments reported in Laptev et al. [2008] and those reported in Ulges et al. [2009] indicate that although the noisy training samples are still able to produce classifiers above the chance level, their performance is significantly poorer when compared to classifiers trained with cleanly labeled training samples.

In our proposal, the issue of lack of training samples becomes even more important, since the framework relies on the ability to train several independent concept detectors to describe actions context. The proposed solution to this issue makes use of TL techniques (Pan and Yang [2009]), in which information contained in auxiliary databases is used to overcome the lack of training samples for the main learning task. Such an approach defies the assumption of most ML algorithms, which is that training and test data come from the same probability distribution. However, as pointed out by Wu et al. [2004], the usage of an external database can be seen as a trade-off between variance and bias: the great variance typically found in small training samples is compensated by the introduction of some bias due to the potentially different distribution obeyed by the auxiliary data.

Such bias is implicitly addressed in Ullah et al. [2010], by the collection of samples visually similar to the target database on the Internet. It is worth to point out, though, that the visual similarity is a subjective metric and does not guarantee the same distribution between the bases. In this work, they also use static action detection – to better specify the action position in the video – and people and object detectors to gather information about context. Since those detectors are trained in material collected from the Internet, as in Ulges et al. [2009], they provide noisy annotations.

Wu et al. [2004] proposed a solution to overcome the lack of training data for leaf image classification by using as an auxiliary external database images from an herbarium (the main database are images of individual fresh leaves, while in the herbarium the leaves are dried and can be tied together in groups). They point out that auxiliary data can be considered both in the definition of the objective function (training phase) and in the class computation for new examples (classification phase). In both cases, the role of the auxiliary data is weighted in relation to the main data, in order to reduce the bias introduced by the external database.

Another example of transfer learning in Computer Vision – yet in still images – is provided by Kumar et al. [2009], for face recognition. Kumar et al. [2009] observe that

[1] http://www.youtube.com

examples of generic faces traits (e.g., lips, moustache, curly hair) are easier to collect than examples of every specific people in the database. So, instead of training a face classifier based on face examples, they train 65 classifiers on selected traits and use the scores provided by such classifiers to compute the feature vectors. Our approach is in some ways similar to theirs, if one observes that instead of "face parts" we build classifiers of "context parts" to achieve the main task at hand.

4.2 Proposed Approach

This section details the framework we propose for action recognition. Firstly, the working hypotheses underpinning the proposed solution and their justifications are summarized:

Hypotheses

- **Contextual information is relevant for action recognition**: as detailed in Section 4.1, this hypothesis is suggested by the recent results of Ullah et al. [2010], Marszalek et al. [2009], Sun et al. [2009a] and Ebadollahi et al. [2006b] and indirectly by Wang et al. [2009] and Fei-Fei and Perona [2005] (this last one in scene recognition).
- **Context is best described by high-level concepts appearing in the scene**: this assumption is based both in the work of Ullah et al. [2010] – which uses concept and people detectors – and Marszalek et al. [2009] – in which the correlation between entire scenes and actions is explored. More generally, this assumption is also supported by the broad review of Snoek and Worring [2008], in which descriptors based on several concept detectors are pointed out as the state-of-the-art solution for semantic video retrieval.
- **Transfer learning can be used to obtain useful concept detectors trained on an external auxiliary database**: such hypothesis is generically supported by the TL theory (Pan and Yang [2009]). Additionally, the approaches presented by Ullah et al. [2010], Wu et al. [2004] and Kumar et al. [2009] indicate the viability of applying transfer learning for Computer Vision tasks.

Finally, this entire work assumes that **BoVFs are appropriate representations for realistic videos and still images**: this assumption is supported by our own survey on the field, presented in Chapter 3, and by the several results on BoVF-based approaches summarized there.

4.3 Transferring Concepts from Still Images to Videos

Before presenting the proposed framework itself, some definitions are stated as follows:

Target Database: The target database is the database in which the main task is defined. In our case, the target database is comprised of videos annotated for actions that we want to be able to automatically recognize.

Source Database: The source (or auxiliary) database is an external database, i.e., a database that is nor related to the target database not to the main task. It can be comprised of images and/or videos annotated for visual concepts of interest. In our case, we are interested in any kind of concept which could provide information about the context in the target database. The main characteristics expected for the source database are: a) to be composed of unconstrained images, and b) to have a large enough number of samples for each concept in order to train reliable classifiers for them.

Figure 4.1 illustrates the complete feature extraction process, including information both from the source database – in our case Caltech256 concept images database – and from the target database – action videos Hollywood2 database – until the *contextualized* representation for the video is achieved, that can be submitted to a final SVM classifier.

The process is performed as follows: in (a), the source database is represented. It is composed of images or videos annotated for several different concepts. To build a SVM model for one concept, a negative sample is created by randomly choosing items from the other concepts. That random sampling equals the number of existing positive ones. This processes of creating negative samples is repeated m times, with the aim at creating m *different models for each single concept*. The experiments that will be described in Section 5.1 show that, at least in our experimental setup, to some transfer to occur, m needs to be greater than one. In other words, one single SVM model per concept is not enough to transfer information from Caltech256 to Hollywood2.

In (b) static concept models are built for each negative sample, using a pyramid kernel with two levels (Lazebnik et al. [2006]). That kernel brings the advantage of including some structural information to the BoVF representation.

Meanwhile, (c) shows that a video summary is created according to a simplified version of the algorithm presented in de Avila et al. [2011]. Such a summary greatly

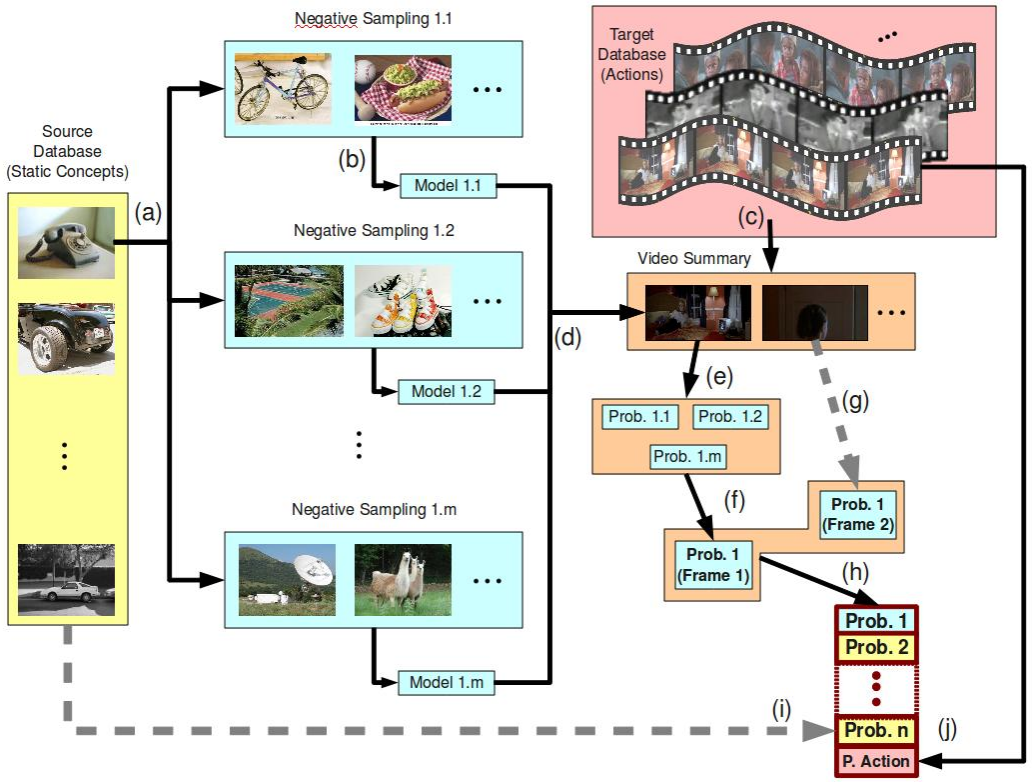


Figure 4.1. Overview of the feature extraction process for the proposed transfer framework. Firstly, several models are created from an auxiliary (source) database (a–b), and applied to the target (action) database (c–d). The results of those models are combined in several ways (e, f, g) and then used as features for the final action classifier (h), together with the results of the baseline classifier (j). More details will be provided in the text.

reduces the amount of redundant frames to be processed, making the overall processing more efficient with little loss of information. As an example, most summaries for Hollywood2 video fragments – which are about a few seconds long, at a 30 frames/second rate – ended up composed of two to four distinct frames. A visual exam of the summaries has shown that the frames selected by the summarization algorithm are really distinct from each other. This means that each summary frame is indeed a source of new information, instead of a source of redundancy.

By the other side, the usage of summary frames instead of a unique keyframe comes from the assumption that a larger number of frames can potentially increase the amount of information about the context for the final action classifier, and avoid some misclassifications due to a "badly" selected keyframe. Figure 4.2 illustrates different frames collected from a unique video segment from the Hollywood2 database (Marszalek et al. [2009]) to illustrate some cases in which summary frames can potentially

increase the probability of a correct classification.

Proceeding with Figure 4.1, in step (d), the m models created in (b) are applied to the summary frames. This process generates then m *different probabilities for each concept* (e).

At this point, an issue of how to combine these results to decide if the concept is or not present in the frame arises. In this work, we tested three different combination strategies: majority voting, average probability and most probable result (those strategies will be detailed in Section 4.3.1). In any case, once a combination strategy is selected, we end up with *a unique decision about that concept for the frame* – step (f).

The arrow labeled (g) indicates the repetition of the process described from (a) to (f) for every frame in the summary. This means that each frame will have a different probability of occurrence for the concept attached to it. Since the summaries are not the same size (in terms of number of frames), again we need an approach to select among those results to have a unique decision about the presence of the concept for that entire video segment. This is a constraint imposed by SVM, which only accepts as input feature vectors having the same size.

In that case, it is assumed that the most important information is the *presence or absence of the concept at any point in the video.* Therefore, the most probable result among the frame results is chosen as the final decision for the concept occurrence at a video level – step (h).

The arrow in step (i) indicates the repetition of the process described from (a) to (h) to every selected concept in the database. In other words, each video of the target database is going to have a probability of occurrence of every selected concept of source database. This stack composed of n concept probabilities is added to the probability provided by the baseline action classifier in step (j), to compose the feature vector that will represent the video for the final SVM action classifier – step (k). That baseline probability of step (j) is computed from a classical SVM action classifier trained on STIP-based BoVFs.

4.3.1 Strategies for Combining Classifiers

In the process depicted in Figure 4.1, step (f), it has been seen that some form of combination strategy was needed to deal with the m results coming from the m different models generated for each concept. Indeed, the idea of having more than one model for each unique concept came from a set of preliminary experiments, which indicated

Clip: *actioncliptest00086.avi* – ***Action:*** *Running*

Clip: *actioncliptrain00003.avi* – ***Action:*** *Getting Out of Car*

Clip: *actioncliptest00013.avi* – ***Action:*** *Fighting*

Clip: *actioncliptest00030.avi* – ***Action:*** *Hugging Person*

Clip: *actioncliptrain00063.avi* – ***Action:*** *Eating*

Figure 4.2. Illustration of how different frames of the same action clip in Hollywood2 can provide more or less contextual information for the final action classifier.

that a unique classifier would not be able to successfully transfer information from the source to the target database, at least in our setup.

Three different combination strategies were tested and compared in our experiments:

Majority Voting

At its name implies, in majority voting each model result is considered as a 'vote' for the presence or absence of the concept in a frame, and the result with more votes is considered the right one.

Average Probability

In the average probability combining strategy, the decision about the presence or not of the concept is based on the average value of the probabilities output by each SVM model when applied to the frame.

Most Probable Result (or greatest distance form $p = 0.5$)

The farthest the output probability for the concept is from the middle of the probabilities interval (i.e., from $p = 0.5$), the surest the model is about its result. That is because a probability too close to $p = 0.5$ would mean a point very close to the classifier margin and, therefore, with a high chance of being at the wrong side of the separating hypersurface. So, in this combining strategy, the decision is based on the model which presents the greatest value for $|prob - 0.5|$.

4.4 Concluding Remarks

In this chapter, a new solution to human action recognition is presented. This solution is based on two main hypotheses of this thesis:

- High-level contextual information can increase action recognition rates.
- Contextual information can be obtained from auxiliary data sources, cleanly annotated, by means of Transfer Learning.

Finally, the proposed solution is depicted in Figure 4.1 and detailed in Section 4.2. In next Chapter, the experiments performed using the depicted solution are described and have their results discussed in the light of the thesis hypotheses.

Chapter 5

Main Results

The bibliographical review presented in Chapter 3 constituted our first contribution, since it proposes a new framework to approaches for action recognition in videos. Besides, we underwent some experimental explorations of BoVF representations in order to achieve a better foundation to the thesis itself. Such preliminary explorations are detailed in Chapter 6. In this chapter, the main results of this work are described and discussed in detail.

5.1 Experimental Setup

This section describes the experiments which support the main contributions of this work, namely, showing the viability of using Transfer Learning (TL) to overcome the lack of training data in action databases and to provide high-level contextual clues which indeed are able to improve action recognition. Such experiments are based on the proposed solution described in Section 4.2.

Two sets of experiments were conducted in order to evaluate our assumptions. Table 5.1 indicates the instantiation of the methodology proposed in Section 5.1 and illustrated in Figure 4.1. From this instantiation, the experiments were designed to answer the following specific questions:

- Is it possible to transfer knowledge from Caltech256 to Hollywood2?
- Does detecting Caltech256 concepts on Hollywood2 frames enhance action recognition?

Table 5.1. Source and target databases used in the experiments

Generic Element	Instantiation
Source database	Caltech256 photos (Griffin et al. [2007]), described by Hue-SIFT (van de Sande et al. [2008]) binary BoVFs. To introduce some geometric information to the representation, a two-level pyramid kernel (Lazebnik et al. [2006]) was applied. The first experiments use *rotary-phone* examples only and the concepts used in the second row of experiments are: *car-tire*, *car-side*, *rotary-phone*, *telephone-box*.
Target database	Hollywood2 videos Marszalek et al. [2009]), described by STIP binary BoVFs, classified by a multilevel kernel as that of Laptev et al. [2008]. In the first set of experiments, only frames from AnswerPhone action (against all) were tested, using a classifier trained on *rotary-phone* concept from Caltech256. In the second set of experiments, all actions are tested.

Transferring knowledge from Caltech256 to Hollywood2

Firstly, we needed to assess whether image-based classifiers would be able to identify the same concepts in videos frames from an unrelated collection. To establish this, Caltech256 *rotary-phone* images where used to train five ($m = 5$) classifiers with different negative samples. In addition, to verify the idea of representation transfer (Section 2.4), the visual vocabulary was randomly selected from: (a) the source-only; (b) the target only and, (c) a mixed selection of visual words from both source and target samples.

Another evaluation was performed at a kernel level, and the following configurations were tested: linear, χ^2, multilevel Laptev et al. [2008] and pyramidal kernel Lazebnik et al. [2006], all using both one and two-level representations.

Such classifiers were applied in a phone baseline database built on frames of Hollywood2 actions database. That baseline was built by manually collecting 90 positive examples of images presenting phones from the summaries of the Hollywood2 action class *AnswerPhone*. The same amount of negative examples were randomly chosen among frames coming from summaries of videos of all other action classes. The summarization algorithm is a simplified version of that presented in de Avila et al. [2011]. Hollywood2 original training and test sets separation – whose samples come from different movies – was maintained.

Our aim in that first round of experiments is to verify the viability of transfer from Caltech256 to Hollywood2 in the concept level, showing how a transfer classifier would compare with a *classical* one – the baseline. It is interesting to notice that

this baseline provides a *ceil* to the transfer result (instead of a *floor*), given that – from the point of view of traditional Machine Learning (ML) theory – the non-transfer setup, with training and testing coming from the same distribution is the *ideal* one. In other words, the transfer classifier is supposed to work, at best, slightly worse than the baseline. Thus, the main advantage of using a transfer-based classifier is that it opens the possibility of using any additional sources of information which would otherwise be inaccessible to a traditional classifier. Hopefully, such additional information would compensate the bias introduced by the extraneous source dataset.

Using Caltech256 concepts on Hollywood2 frames to enhance action recognition

In this second round of experiments, the feature vectors extracted by the procedure described in Figure 4.1 for every video are submitted to a new SVM classifier, this time with a kernel based on χ^2 distances (in practice, the same of Laptev et al. [2008], but using a unique channel).

Similarly to the case with the concepts, not all the training set were used at once for training the baseline. Instead, all the positive samples for each action were taken together with a random selection of negative samples of the same size. This more lightweight experimental setup (when compared to a typical full Hollywood2 classification setup as those in Laptev et al. [2008] and in Marszalek et al. [2009] was chosen to speed-up the verification of our main assumption: the ability of Caltech256 concept classifiers to enhance action classification on Hollywood2.

5.2 Results and Discussion

5.2.1 Transferring from Caltech256 images to Hollywood2 frames

In this section we show that, indeed, knowledge transfer from Caltech256 to Hollywood2 frames is feasible within the proposed framework (Figure 4.1).

Figure 5.1(a) shows the individual results for Phone classification in Hollywood2 frames, with and without transfer. As suggested by Pan and Yang [2009], such "brute force" transfer led to negative transfer for individual classifiers. Nevertheless, in Figure 5.1(a) the individual results for transfer settings are combined, using five replications per kernel configuration, indicating that in some cases it is possible to achieve results above chance level, although skewed to the positive side. It is possible to observe

also that the less skewed results tend to be provided mostly by the more sophisticated kernels, namely those proposed in Laptev et al. [2008] and Lazebnik et al. [2006]. Figure 5.1(a) shows the same data colored by combination scheme, showing that average and maximum distance from the mid-point probability are those responsible for the best transfer results.

Replications of these experiments with $k = 4000$, indicated no expressive enhancement over $k = 400$, while the results obtained with $k = 100$ individual tend to present stronger fluctuations, thus becoming less reliable.

Figure 5.2 shows the F1 score between the baseline and the results with transfer from Caltech256, using three different sources for the visual vocabulary. Each graph illustrates a different combination scheme for the classifiers. In Figure 5.2(a) individual classification results are combined by majority voting for each item; In Figure 5.2(b), the probabilities of positive samples provided by the classifiers are averaged. In Figure 5.2(c) the combined classifier relies on the individual classifier whose probability is farthest from the middle-point ($p = 0.5$), meaning that in this scheme the final classifier uses the result of the classifier which has the greatest certainty of its response for that item.

Dark blue bars represent the baseline results (without transfer) while the other colors indicate different visual vocabulary sources. Each group of bars was computed using a different kernel.

In terms of combining schemes, these graphs show that: a) majority voting seems to be appropriate only for traditional classification setups, while being unstable in transfer setups; b) using the average probability for decision is better for transfer setups than to a traditional one and, c) the usage of the greater certainty classifier seems to be the most stable between transfer and traditional setups.

In addition, these graphs reinforce that the differences among kernel configurations or vocabulary sources are not statistically significant. Such insensibility to vocabulary sources is an important result for our transfer setup, since it means that there is no transfer of representation knowledge and positive transfer results can be obtained simply by a combination of "weak" classifiers applied in a "brute-force" transfer fashion.

5.2.2 Using transfer to improve action recognition

Based on the results of the previous experiments, it is now assumed that it is viable to transfer at least some knowledge obtained from the source database (Caltech256) to the frames of the target database (Hollywood2) at a per concept basis. The results

of this new set of experiments are shown in Figure 5.3, which presents the differences between precision values of the transfer results in relation to the no-transfer baseline (STIP-only) showing an increase in the majority of precision values when transfer is applied.

Table 5.2 shows the differences for the average combination scheme and their statistical significance. From Figure 5.3 and Table 5.2 it is possible to see that there are 4 decreases in precision only, out of 12, and 3 of them are **not significant**. For the the precision increases (8 out of 12), all but *FightPerson* and *SitUP* results are **significant**. In other words, 6 out of 8 increases in precision are statistically significant, meaning that at least for half of the available actions in Hollywood2, our solution is able to increase recognition rates.

Observing Figure 5.3 for different actions, it is possible to see that the *DriveCar* action had the largest enhancement, what should be expected, since two transfer concepts in this experiment were related to cars (*car-tire* and *car-side*). However, unexpected negative results come up for *AnswerPhone* and *GetOutCar*, since the transfer concepts were semantically related to them either.

In case of *AnswerPhone*, the insignificant impact of the phone concept can be explained by the fact that telephones usually appear as small objects in the scene, therefore playing a small part in the context. In case of *GetOutCar*, we notice that this action is notoriously tricky in Hollywood2, since there are scenes in which a car not even appears in the scene, or the action is extremely occluded, as can be seen in Figure 5.4.

The positive results of the other classes (*HandShake*, *Kiss*, *Run*, *SitDown* and *StandUp*) could also be considered somewhat unexpected, given the apparent *unrelatedness* of them with the selected transfer concepts. Such results reinforces our main thesis that general concepts can have a positive impact on the overall action recognition performance. In addition, it also reinforces our hypothesis that even apparently unrelated concepts are able to convey indirect contextual clues to the action classifier.

Graphs in Figure 5.5 show the Receiving Operator Characteristic (ROC) points for the class *AnswerPhone* (a bad result), for ten different replications (changing the random validation set) and for the three cases of source classifier combinations (average probability, majority voting and maximum distance from the middle-point probability). In this particular case, transfer information seems not to have any influence on the classifier.

Figure 5.6 shows the the ROC points for the *DriveCar* action, in which the gains in precision were the largest ones.

Table 5.2. Differences in precision values with their statistical significance at a 95% level

Action	Difference	Significant?
AnswerPhone	-0.0132	no
DriveCar	0.0395	**YES**
Eat	-0.0152	**YES**
FightPerson	0.0060	no
GetOutCar	-0.0036	no
HandShake	0.0087	**YES**
HugPerson	-0.0009	no
Kiss	0.0181	**YES**
Run	0.0090	**YES**
SitDown	0.0283	**YES**
SitUp	-0.0073	no
StandUp	0.0390	**YES**

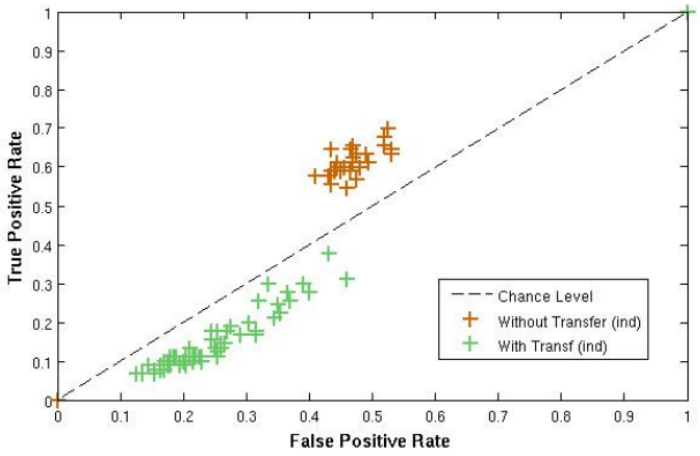
Finally, the same graphs for *GetOutCar* action (Figure 5.7) show another interesting effect of using transfer information, which was found in other actions either (not shown to avoid redundancy). In this case, although the average precision gain is very small or negative (depending on the combination scheme), it is possible to see how the transfer-based classifiers tend to be less biased than the baseline classifier.

5.3 Concluding Remarks

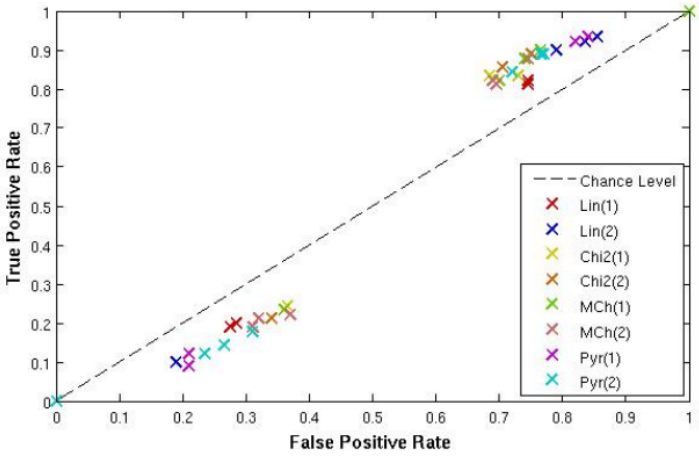
The set of experiments described in this chapter was aimed at assessing the hypothesis that Transfer Learning (TL) makes it possible to use the information contained in already cleanly annotated databases to improve action recognition results in challenging, realistic videos. In a first step, an image-to-frame transfer evaluation was performed using the Caltech256 as the source concept database and Hollywood2 as the action target database, indicating that the transfer of knowledge about concepts is possible at frame level, despite the different distributions between source and target databases.

The second round of experiments is built on the insights obtained in that first round, we selected four concepts from the source database Caltech256 and applied the proposed solution. The results show that most differences in action recognition with transfer are positive and statistically significant, when a suitable combination of classifiers trained with source examples is employed.

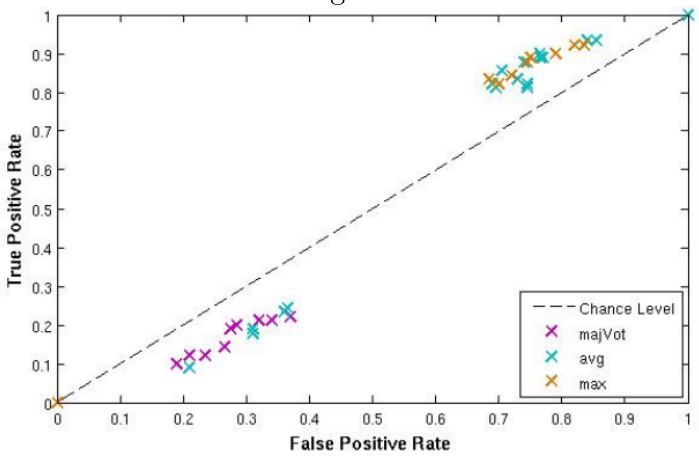
The most important outcome from those experiments is a clear indication that such transfer-learning-based framework can effectively enhance action recognition rates, thus validating our thesis.



(a) Individual Results, showing that initially, there is negative transfer.

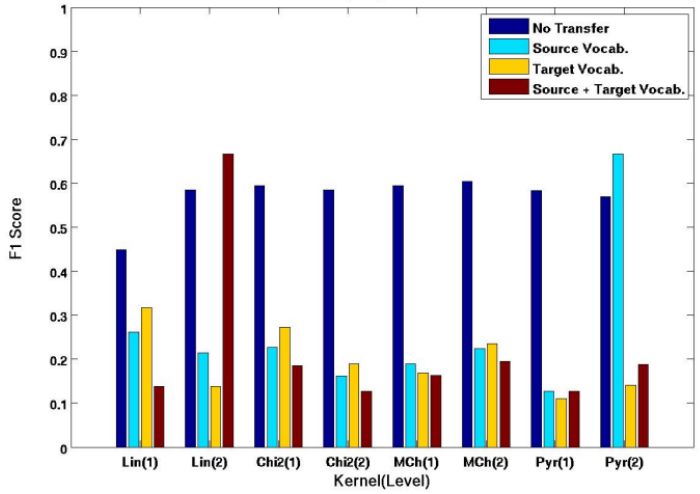


(b) Results combined (colored by kernel), showing that some model combinations can overcome negative transfer.

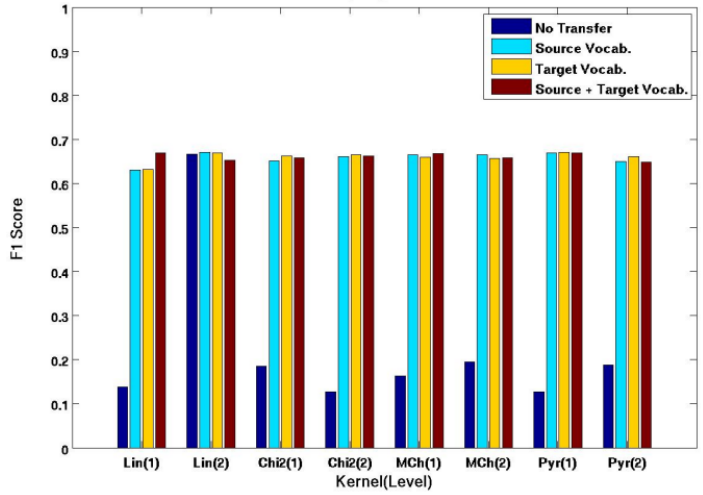


(c) Results Combined (colored by combination strategy), showing that average and maximum distance from the $p = 0.5$ work better than majority voting.

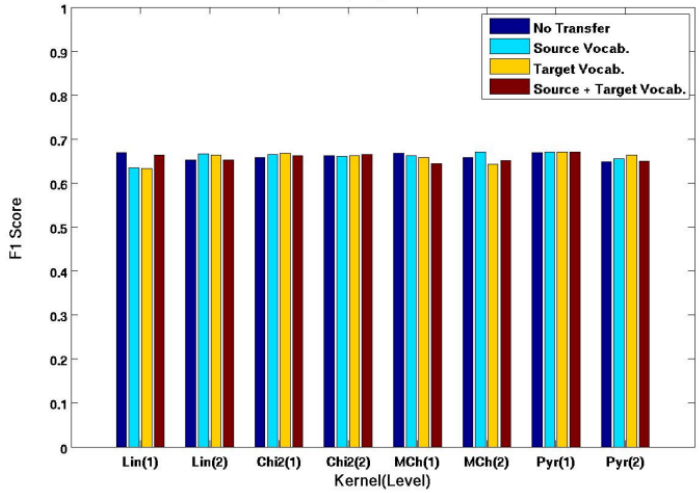
Figure 5.1. Transfer of Knowledge about Phones (k=400, m=5)



(a) Majority Voting is generally better for non-transfer setups.



(b) Average Probability is generally better for transfer setups.



(c) Max Distance from p=0.5 works equally well for both transfer and non-transfer setups.

Figure 5.2. F1 Scores show that kernel and vocabulary sources do not present significant differences for transfer, while the combination scheme can greatly influence transfer results. (k=400, m=5)

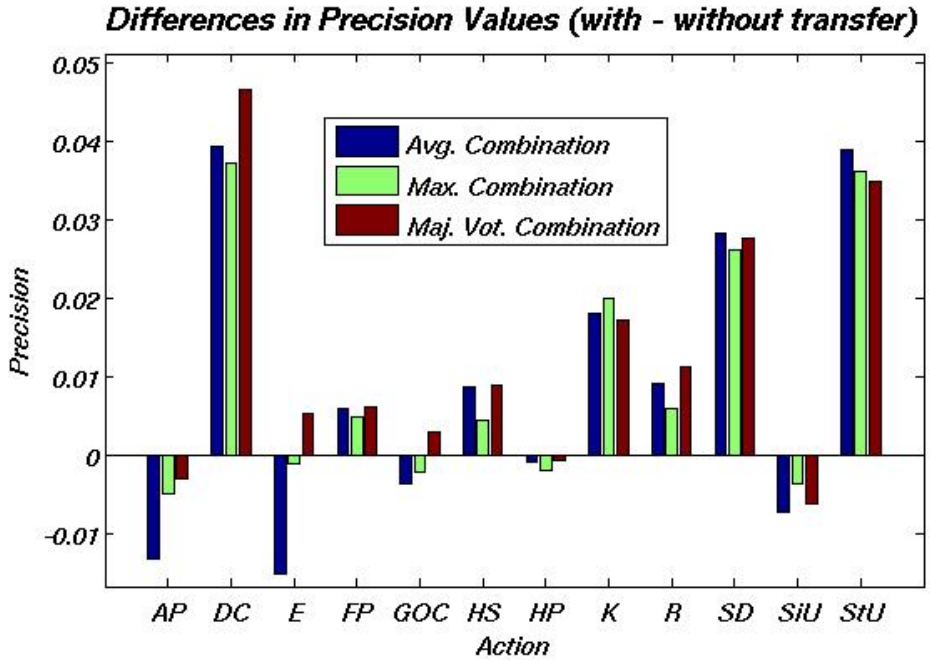
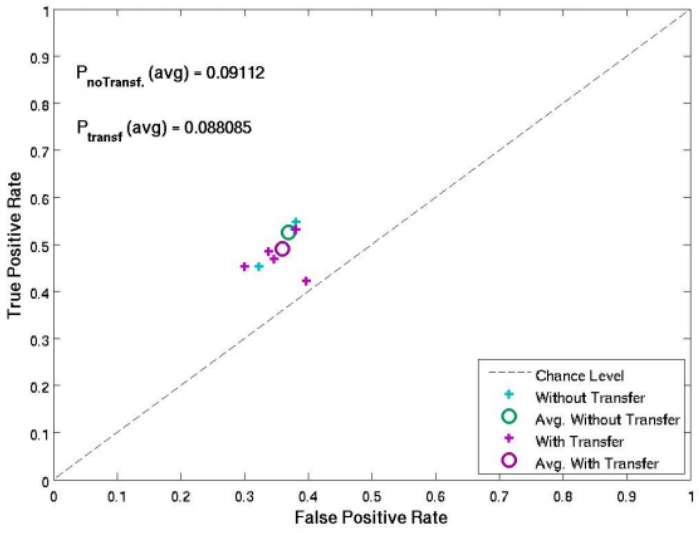
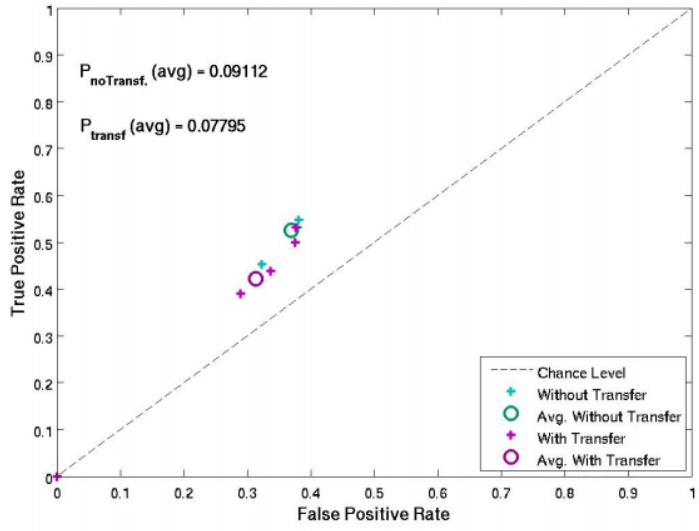


Figure 5.3. Indicates that even with only *4 (four)* concepts used for transfer, ***most precision values increase***, no matter the classifier combination scheme applied for the concept classifiers. Actions: AP – AnswerPhone, DC – DriveCar, E – Eat, FP – FightPerson, GOC – GetOutCar, HS – HandShake, HP – HugPerson, K – Kiss, R – Run, SD – SitDown, SiU – SitUp, StU – StandUp. Transfer Concepts: car-tire, car-side, rotary-phone and telephone-box.

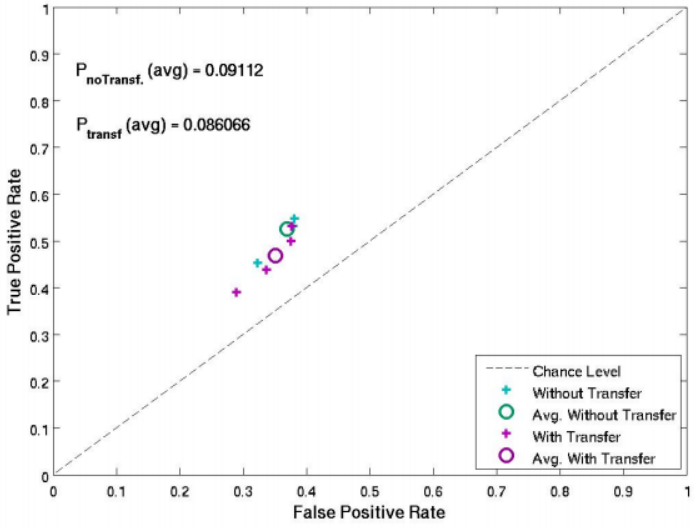
Figure 5.4. A sequence exemplifying a difficult example of *GetOutCar* action. Observe that the car is presented in a frontal point-of-view (instead of *car-side* and the visual information related to the action of getting out of the car is very small (magenta arrow).



(a) Majority Voting

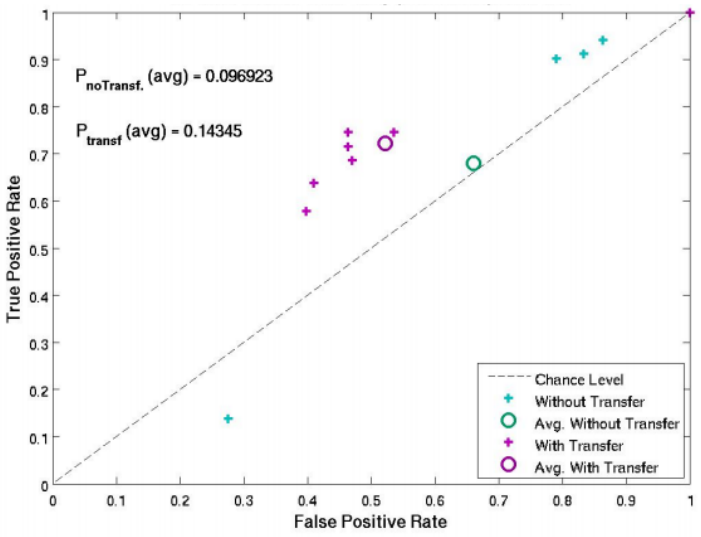


(b) Average Probability

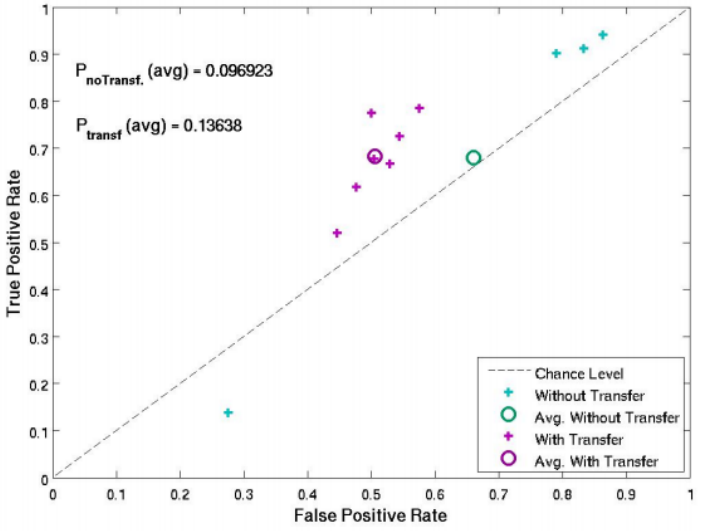


(c) Max Distance from p=0.5.

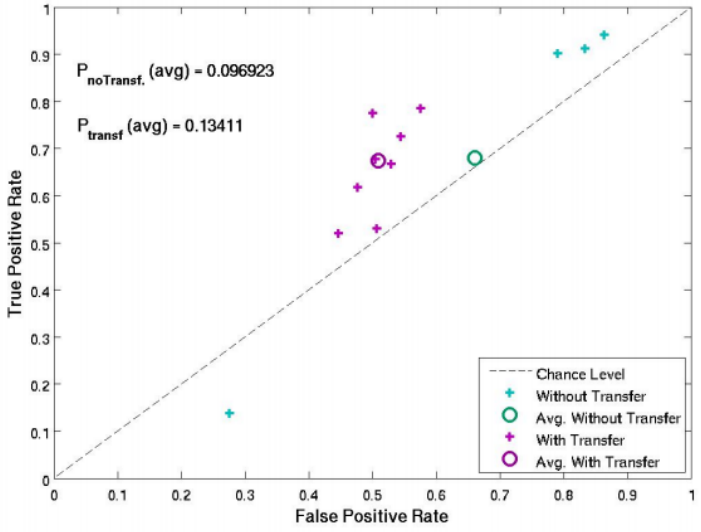
Figure 5.5. *AnswerPhone* transfer results (k=400, m=10), showing a case in which transfer is not able to have an influence on classification results.



(a) Majority Voting

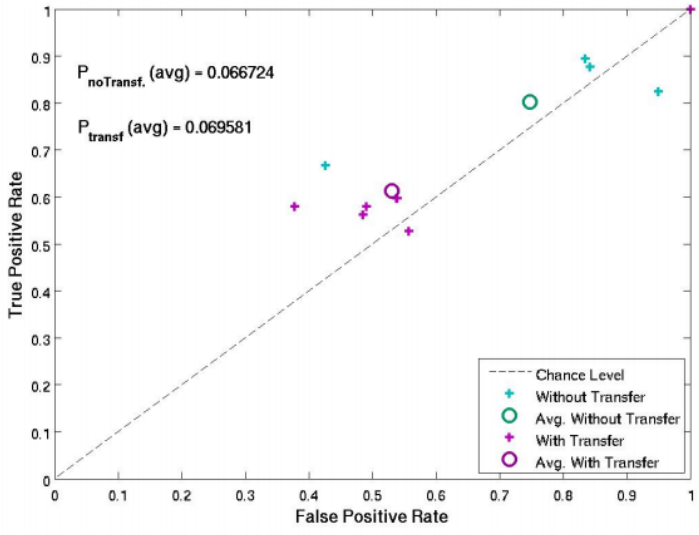


(b) Average Probability

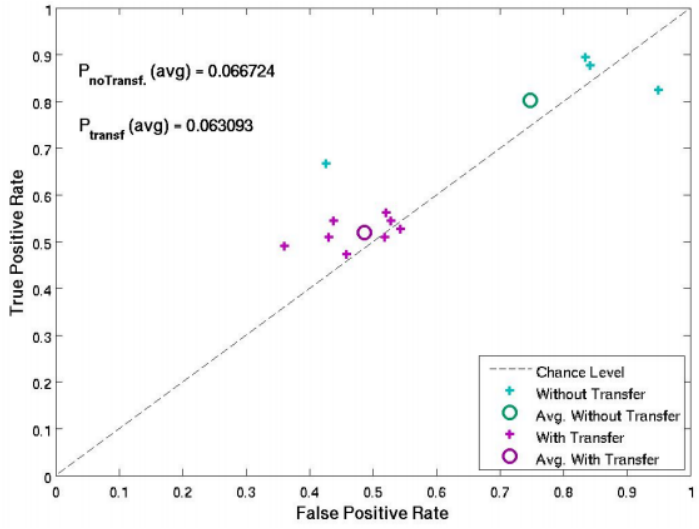


(c) Max Distance from p=0.5.

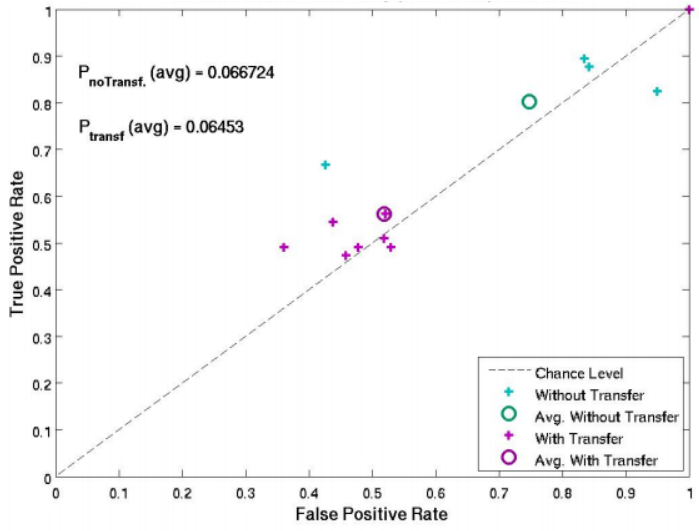
Figure 5.6. *DriveCar* (best case) Transfer Results (k=400, m=10) show how the information introduced by transfer move the average precision away from the chance-level line.



(a) Majority Voting



(b) Average Probability



(c) Max Distance from p=0.5.

Figure 5.7. *GetOutCar* transfer results (k=400, m=10) is an example (among others) in which even when transfer does not move the results away from the chance-level line, they tend to be more unbiased with transfer.

Chapter 6

Additional Results

6.1 On Vocabulary Building

Despite the widespread usage of k-means algorithm in visual vocabulary building (see Section 2.2), it is not clear in the literature whether this choice is worth its cost. K-means computational complexity is considerably high, therefore imposing scalability constraints on BoVF-based methods. Our own results, like in Lopes et al. [2009d], for instance, indicate a random oscillation in the action recognition rates obtained with k-means. Such results arose the supposition that a random selection of visual words could be as good as k-means, or at least good enough with a much lower computational cost.

With this in mind, we performed a series of experiments on Weizmann database comparing three strategies for vocabulary building:

1. k-means (the baseline);

2. Pure Random Selection (PRS) and;

3. a random selection followed by a greedy strategy aimed at enhancing the result of PRS, which we called Enhanced Random Selection (ERS).

Enhanced Random Selection (ERS)

The ERS strategy is based on the observation found in Jurie and Triggs [2005], that k-means has a tendency to select too many points in feature space regions where the distribution is denser. This can produce a vocabulary with a great amount of very similar words. The solution proposed by Jurie and Triggs [2005], though, involves a

new clustering algorithm and dense point sampling, compromising the goal of reducing the overall computational complexity.

ERS has a random subset of points as input. In a greedy strategy, the algorithm takes every point of that subset and eliminate from the vocabulary any other point whose distance to the current one is below a certain threshold. This strategy leads to a better distributed visual vocabulary, as can be seen in Figure 6.1.

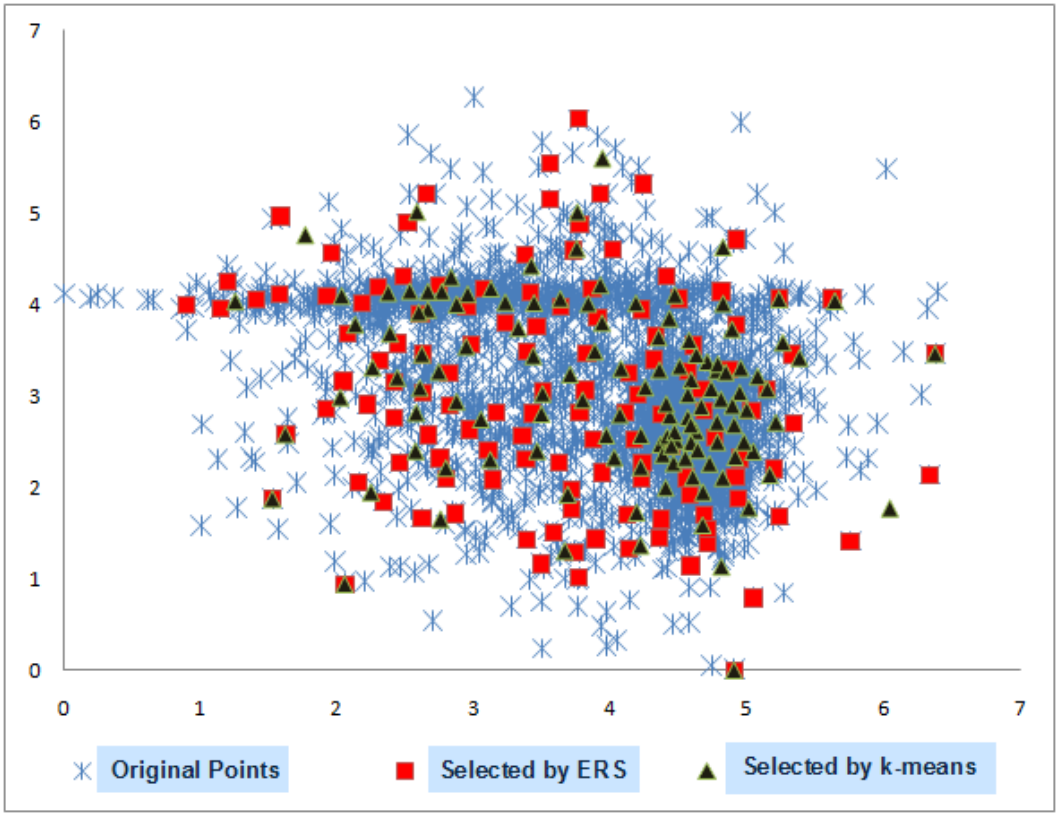


Figure 6.1. Comparing vocabulary centroids selected by k-means and ERS in a 2-dimensional feature space. The original points are produced by a summation of two gaussian distributions with distinct parameters. It is possibl to see that k-means selection follows the original distribution more closely, while using ERS method, one can obtain a set of points better distributed in space.

6.1.1 Results and Discussion

The evaluation of the three proposed strategies for visual vocabulary selection was performed on the Weizmann database, presented in Gorelick et al. [2007].

The Weizmann actions database is comprised of short video segments, containing nine different people performing ten different actions. The actions considered are *bending to the floor, jumping-jacking, jumping forward, jumping in place, running,*

walking laterally, jumping on one foot, walking, waving with one hand and *with two hands*. Snapshots of the Weizmann database can be seen in Figure 6.6. This database has been used for several authors and became a the first *de facto* standard for human action recognition evaluation.

Recognition rates were obtained by SVM classifiers. A five-fold cross-validation was applied to estimate the margin error parameter of SVM, varying its value from 10^{-8} to 10^{8} in multiplicative steps of 10. Finally, each experiment was repeated ten times for each algorithm and vocabulary size.

Vocabulary sizes were varied from 100 to 1500 approximately[1], with and without PCA – except for PRS with χ^2 distance, since this distance is not defined for negative values. The idea in this case is to evaluate the impact of PCA in classification.

The graphs in Figure 6.1.1 show the mean recognition rates for all cases. In them, it is possible to see that k-means provides recognition rates a little higher than the other algorithms in most cases. The visual examination also shows that PCA-based solutions (left) tend to produce slightly worse results, probably due to the fact that PCA tends to produce a "flatter" feature space, reducing discrimination ability. Yet, linear kernel tends to be less stable than χ^2. Such difference is best evidenced in Figure 6.1.1, in which the variance of the ten runs is shown for each case.

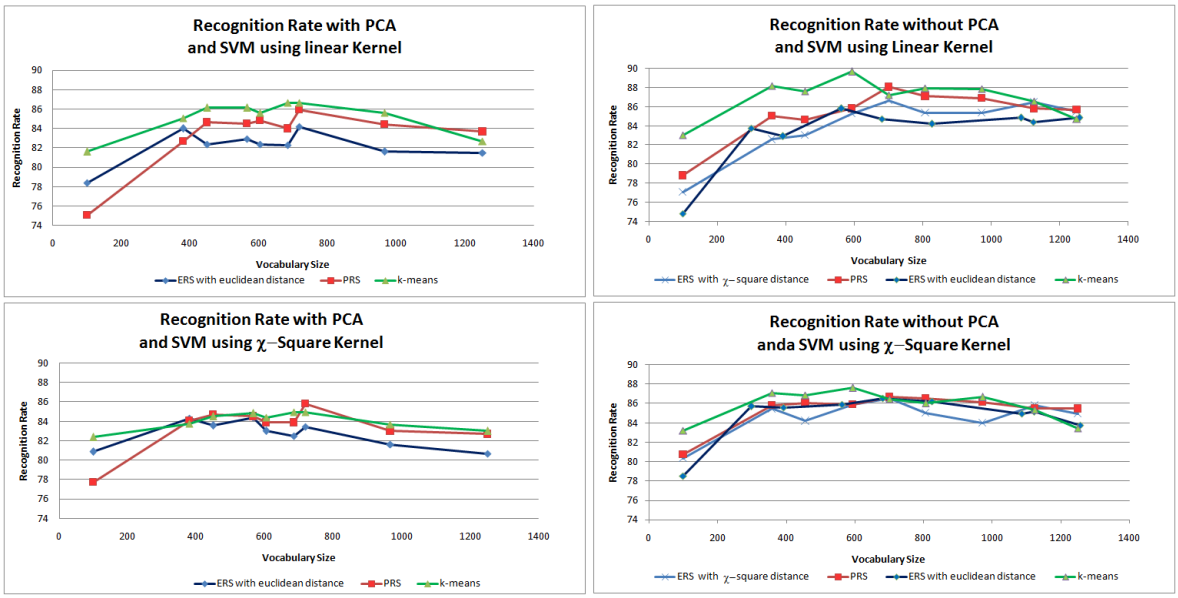


Figure 6.2. Recognition rates obtained by the three algorithms explored, both with and without PCA.

Since the results without PCA present better results, they were chosen to a more detailed analysis. Figure 6.1.1 shows the confidence intervals for the recognition rates

[1]The exact vocabulary size was determined by the results of ERS.

for a vocabulary size around 700, for different algorithms and kernels. That table shows that, even when the confidence is relaxed to 90%, the differences are not statistically relevant.

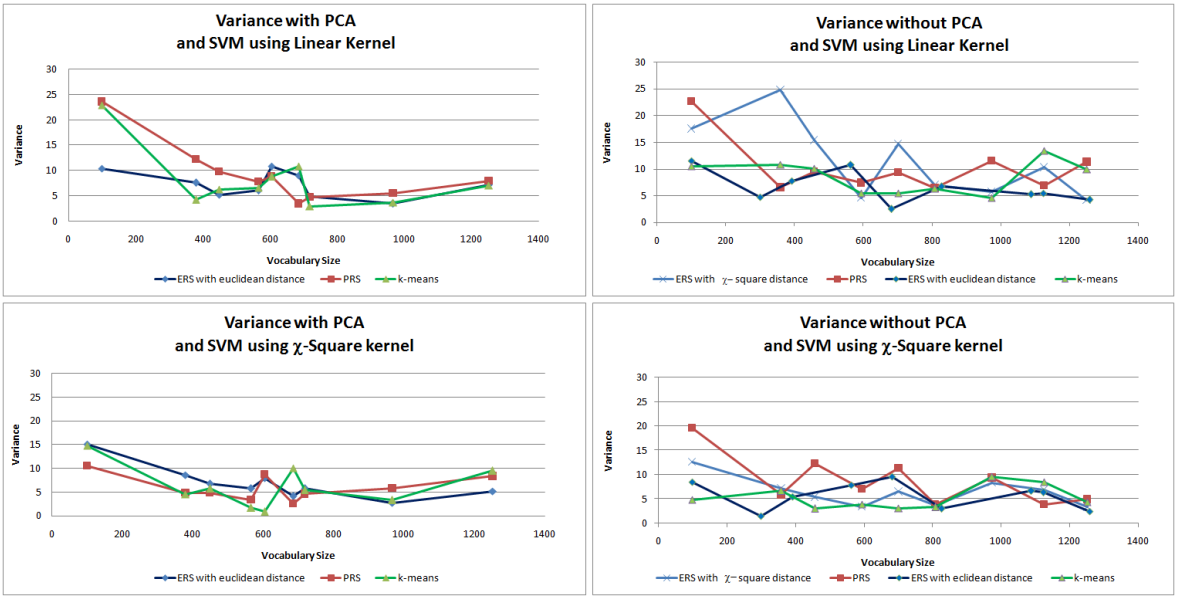


Figure 6.3. Variances for the three algorithms explored, with and without PCA.

Such results provide two important pieces of information: firstly, contrary to the intuition suggested by Figure 6.1 and by Jurie and Triggs [2005], a better distributed vocabulary was not able to significantly enhance video representation over a k-means based selection. Besides, there is no statistical difference between the recognition rates achieved by using k-means and PRS or ERS. However, Figure 6.1.1 also indicates that results based on k-means with χ^2 kernel tend to be more stable, since the confidence interval is shorter than all the other cases. Disregarding this downside of a slightly increased instability in the action classifier, the overall results indicate that it is indeed possible to circumvent k-means by a constant cost PRS selection, without a significant loss in average classification rates. Such conclusion is similar to that of Viitaniemi and Laaksonen [2008], for still images and we consider that the usage of random vocabularies in the experiments described in Chapter 5 is therefore justified.

6.2 Adding Dynamic Information to BoVFs

The simplest extension of BoVF representations for videos can be done by collecting points from every frame in the video segment under consideration and count all them to create the corresponding histogram. The drawback of such a simplistic approach is that

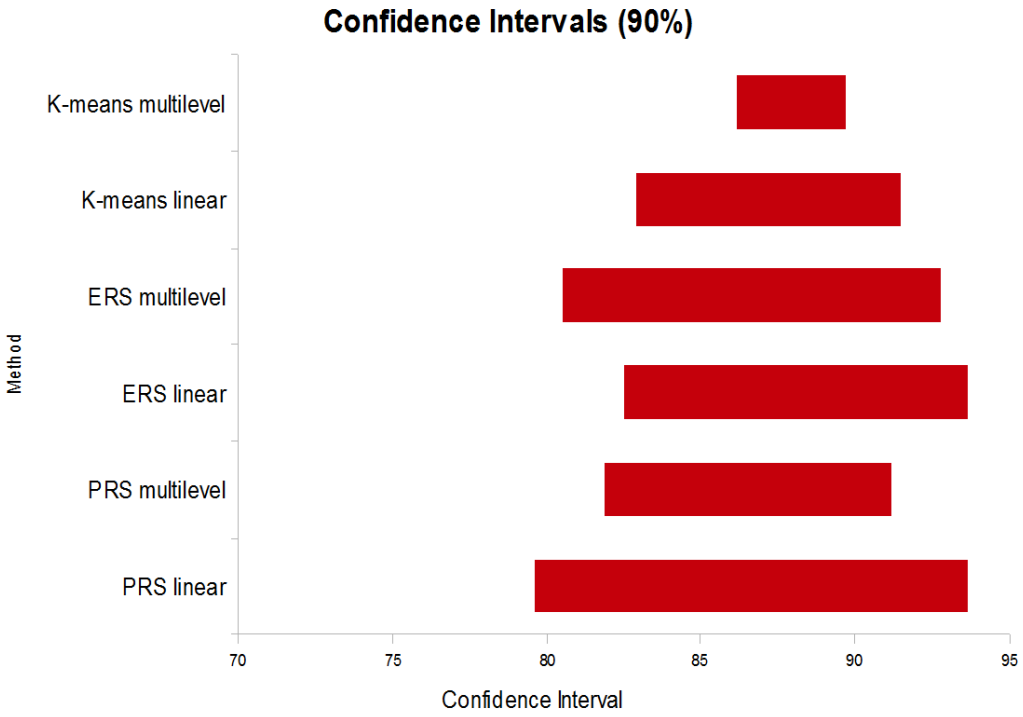


Figure 6.4. Confidence intervals for the recognition rates using $k \approx 700$ and without PCA, with 90% of confidence.

it disregards what happens along the temporal axis, where the dynamic information actually is.

Therefore, for videos to be represented by BoVFs, specially with the goal of distinguishing among human actions, an important issue that arises is how to represent dynamic information. Most existing proposals consider the video as a spatio-temporal volume and then describe *volumetric patches* around 3D interest points (Dollar et al. [2005], Scovanner et al. [2007], Liu et al. [2008b], Laptev et al. [2008] and Niebles et al. [2008]).

However, we envisioned an alternative approach to gather both appearance and dynamic information, which is based on 2D descriptors applied to the spatio-temporal video planes. In other words, we propose that the BoVF representation for videos can be built by collecting 2D interest points, given that such points are selected not only from the traditional frames (xy planes), but also to those ones composed of one spatial dimension and the time axis (which we call spatio-temporal frames). This idea is illustrated in Figure 6.5.

The pure spatial planes (those in the xy direction) are the video original frames. Stacked together, they form a spatio-temporal volume that can be spanned not only in

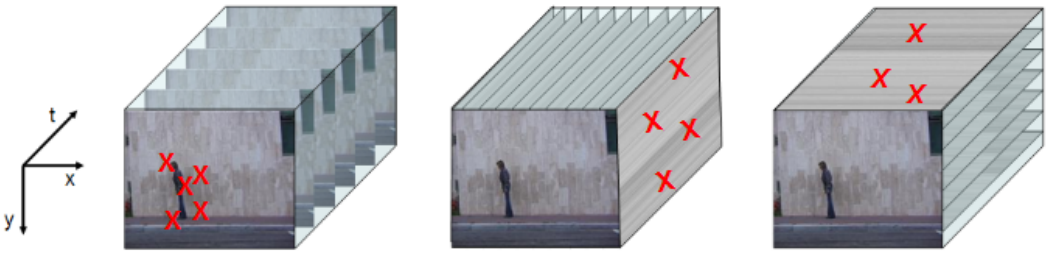


Figure 6.5. Selecting and describing 2D interest points from *spatio-temporal frames*.

the xy direction, but also in xt or yt directions, forming the *spatio-temporal "frames´´*. In other words, the spatio-temporal frames are those planes formed by the temporal axis and one of the spatial axis (xt and yt). The assumption behind this proposal is that such 2D descriptors are able to capture dynamic information from videos without the need for 3D extensions of those descriptors.

6.2.1 Experimental Results

The idea proposed above was tested in the task of human action recognition, by means of a number of experiments performed on the Weizmann actions database (Gorelick et al. [2007]).

Weizmann database is somewhat simplistic, presenting few variations in scale and appearance. In other words, the main differences among actions are due to motion, so the premise for using such a simple database is that a descriptor which better captures dynamic information is also better suited for action recognition on it.

Figure 6.6. Snapshots of the Weizmann Gorelick et al. [2007] human actions database.

The 2D detector/descriptors chosen are those delivered by the well-known SIFT algorithm presented in Lowe [1999] and by a newer competitor SURF, proposed in

Herbert Baya and Gool [2008] (see Section 2.1 for details on them).

The issue of the best combination of frames among the xy (traditional frames), xt and yt (spatio-temporal frames), for both SURF and SIFT, is addressed in the first set of experiments. Then, the best results from those experiments are compared to the result achieved with the state-of-the-art 3D STIP descriptors from Laptev and Lindeberg [2003]. In all cases, an extensive search for the best vocabulary size (the k value of the k-means algorithm) and SVM penalty error parameter (C) is performed. The following subsections describe the experiments performed in more detail.

Do points from spatio-temporal frames really capture video dynamics?

In this set of experiments, interest points coming from the three different sets of 2D frames (xy, xt and yt) are tested separately and in different combinations ($xy + xt$, $xy + yt$, $xt + yt$, $xy + xt + yt$). Both SURF and SIFT-based BoVFs are computed.

For every set of frames and interest point algorithm, the experiments are carried out as follows: the BoVF extraction process is performed on several values for the vocabulary size k. In case of planes combinations, the BoVFs obtained for every plane set are concatenated to form a final BoVF. Values between 60 and 900 for the final BoVF representation are tried, with steps of 60.

For each vocabulary size, an extensive search for the SVM penalty error C that would provide the higher recognition rate was done. A logarithmic scale between 10^{-10} and 10^{10} with 10 as a multiplicative step was used for the C values. Every recognition rate for every k and C was measured on a 5-fold cross validation run. Once the best k and coarse C is found, a finer search of C values is performed, around the previous best one, this time with a multiplicative step of 10^{-1}.

With the best k and best C at hand, ten 5-fold cross validation are run on the database, varying the random selection for the folds. Then, the ten mean recognition rates found at these runs are averaged to compute the confidence intervals.

Figure 6.7 shows the confidence intervals for the best recognition rates achieved with SURF points using different combinations of sets of frames (at a 95% confidence level). The combinations are indicated in the ordinate axis of that graph, while the recognition rates are indicated in the abscissa axis.

From this graph, it is possible to see that just by using information from one of the spatio-temporal frames to compute the video BoVF provides significant improvement on the recognition rate over the BoVF created only from points detected on the original xy frames ($\pm$11% higher).

Also, the combination of the BoVFs coming from the xy (pure spatial) frames

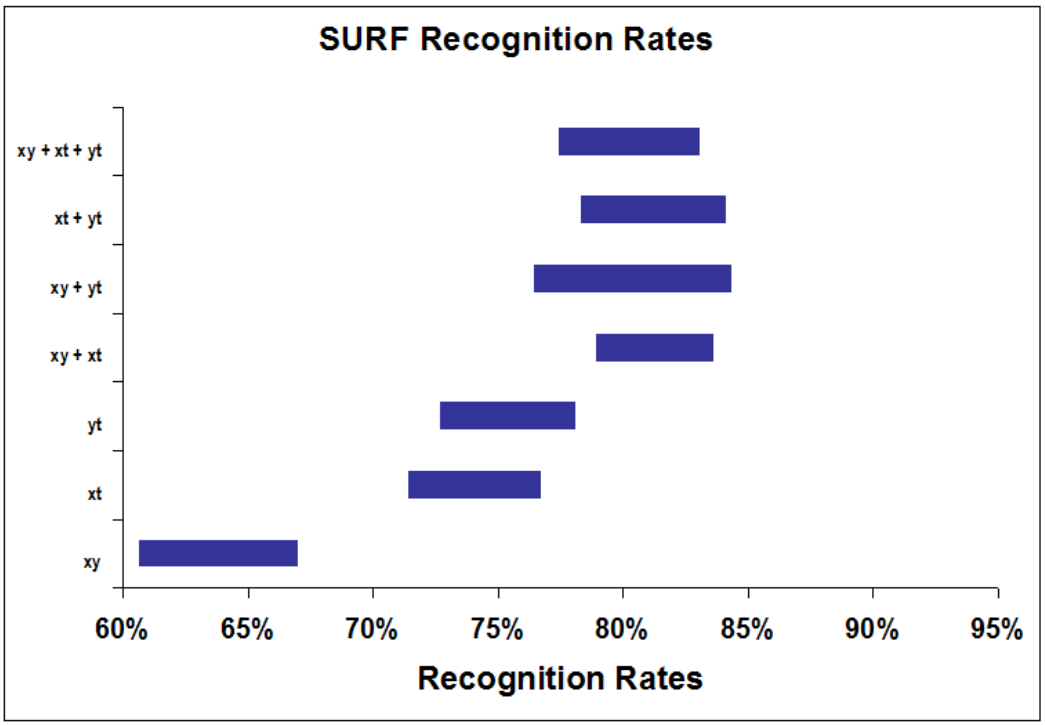


Figure 6.7. Confidence intervals for the recognition rates obtained with SURF points gathered from different frames sets, at a confidence level of 95%.

together with one of those coming from the spatio-temporal frames (xt OR yt) performs even better ($\pm 5\%$ higher than the result achieved with data from spatio-temporal frames only).

Nevertheless, combining the points from all the frames together does not provide further improvement on recognition rate, as it could be expected at a first sight. As can be seen from Figure 6.7, the recognition rates provided by the combinations $xy+xt$, $xy+yt$, $xt+yt$ and $xy+xt+yt$ have no statistically significant difference. This result indicates that while pure spatial and the spatio-temporal frames are complementary between them, spatio-temporal frames from different directions carry redundant information.

Figure 6.8 shows the results of the equivalent experiments with SIFT descriptors. As it can be noticed, these results are quite consistent with the SURF ones, including the fact that the recognition rates achieved by using points from xy and yt frames together are the best ones.

Since the Weizmann database is specifically focused on human actions, these results provide a strong indication that 2D interest points descriptors can indeed be used to capture dynamic information in videos, when applied to the spatio-temporal frames.

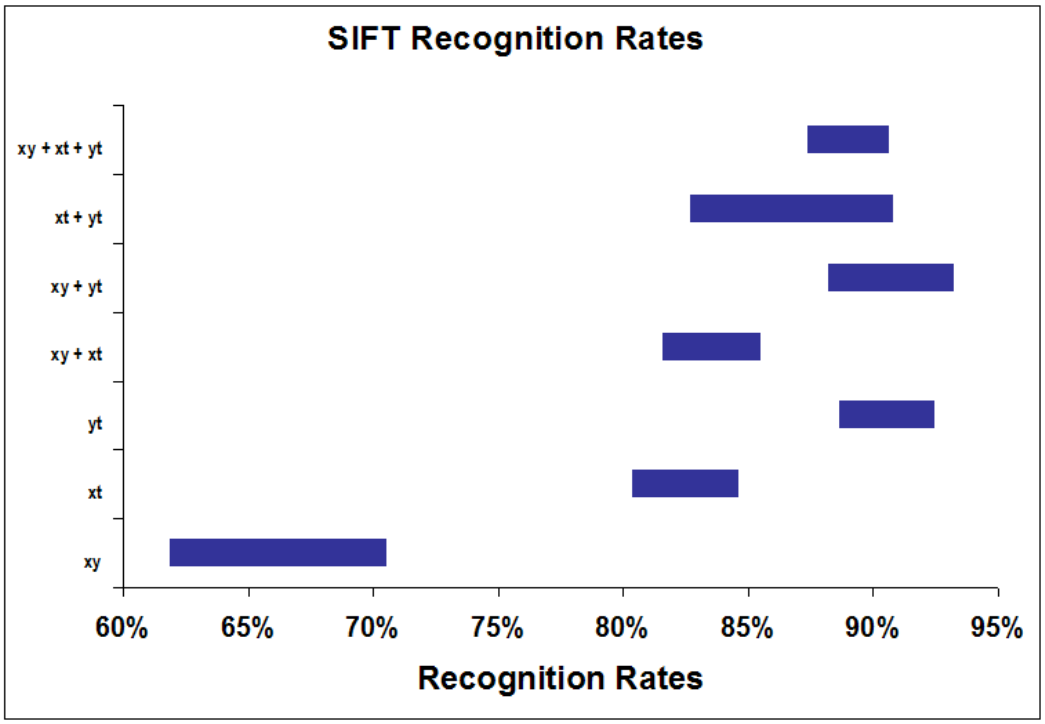


Figure 6.8. Confidence intervals for the recognition rates obtained with SIFT points gathered from different frames sets, at a confidence level of 95%.

Finally, our best results – in both cases obtained by the combination of xy and yt frames– are compared to other published results on the same database in Table 6.1. Of course, this is quite a coarse comparison, because of the lack of a standard experimental protocol. Anyway, assuming confidence intervals similar to our ones, some discussion can be done on these numbers. Dollar's features (brightness gradients on space-time cuboids) are added with geometric information from a constellation model in Niebles and Li [2007], and a 3D extension for SIFT is proposed in Scovanner et al. [2007]. These proposals are the more directly comparable to ours, since they deal with the temporal dimension with extensions of 2D descriptors, without further improvements to the basic BoVF. In those cases, our proposal produces better or equivalent recognition rates when SURF is applied, and much higher ones when SIFT is the choice for point selection and description.

The recognition rates published in Niebles et al. [2008] and Liu et al. [2008b] are considerably higher than our best SURF-based ones and statistically equivalent to our best SIFT-based result. However, it is worth mentioning that in Niebles et al. [2008], a classifier based on pLSA is applied and in Liu et al. [2008b], Dollar's features are fused with features obtained from the actor's spatio-temporal volumes built from body silhouettes. This means that their method looses its model-free nature, since it relies on

Table 6.1. Comparing recognition rates of BoVF-based approaches applied to the Weizmann database. Some details of each comparing approach are provided in the text.

Paper	Rec. Rate
Niebles and Li [2007]	72.8%
Ours (SURF)	$81 \pm 3\%$
Scovanner et al. [2007]	82.6%
Liu et al. [2008b]	89.3%
Niebles et al. [2008]	90%
Ours (SIFT)	$91 \pm 3\%$

the existence of the human body at a certain scale and on the ability to separate body silhouette from the background. By the other side, our plain BoVF implementation – with SIFT as the point selector/descriptor applied both to the traditional xy frames and to the yt spatio-temporal frames concatenated together – provides a similar average recognition rate, even using a simple linear SVM classifier.

The results and analysis of this first set of experiments have been published in Lopes et al. [2009d].

Comparing SURF, SIFT and STIP

The next sequence of experiments is aimed at comparing the best results of SURF and SIFT on spatio-temporal frames with those that could be achieved with the 3D STIP descriptor from Laptev and Lindeberg [2003].

Figure 6.9 shows the results for various vocabulary sizes. As it can be seen, except for small values of k, the spatio-temporal SIFT (using xy and yt frames) based approach is consistently better than the other two, and the spatio-temporal SURF is consistently the worst among the three alternatives.

The superiority of SIFT over SURF is probably due to the fact that SIFT selects more points than SURF, providing a denser sampling than SURF. However the most important conclusion derived from such results is that they contradict the claims of SURF authors for the equivalence of SURF and SIFT, which does not hold, at least in this scenario.

From Figure 6.9 is also possible to see that the best results with the STIP-based BoVF is achieved at a vocabulary size of 540 words, while the best result using

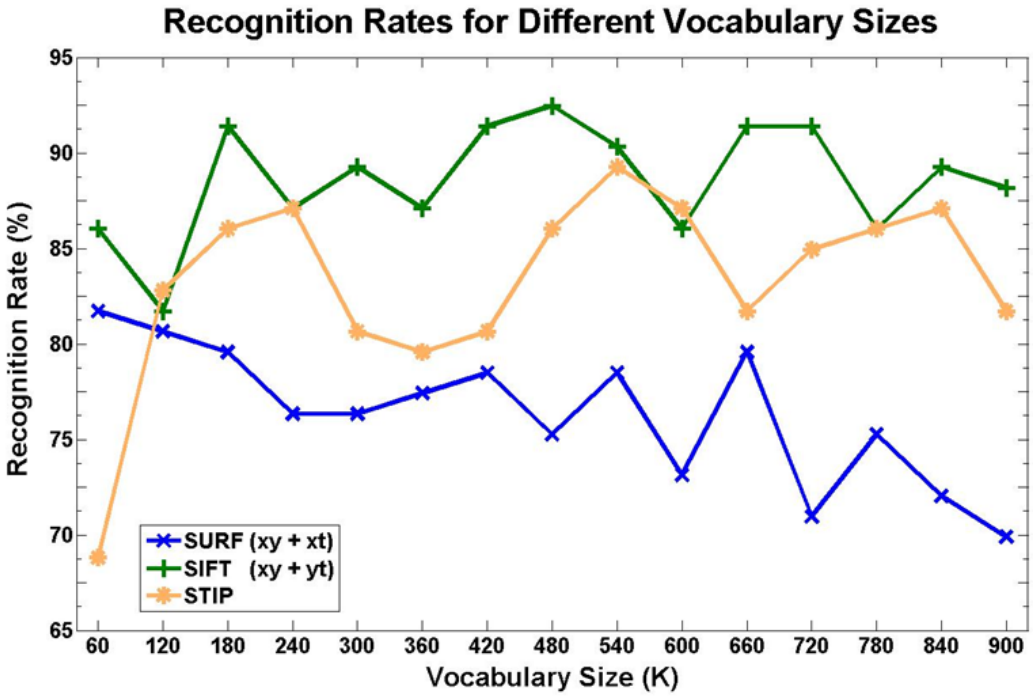


Figure 6.9. Comparing results for spatio-temporal SURF, spatio-temporal SIFT and STIP at various vocabulary sizes.

the spatio-temporal SIFT was around 480 words. The exact recognition rates – at a confidence level of 95% – are 88.5±4% and 90.7±5% – thus showing no statistical difference.

Taking a closer look at Figure 6.9, though, it is possible to see another similar peak for the spatio-temporal SIFT results, at a much smaller vocabulary size of 180.

Figure 6.10 includes the results for both peaks, showing that there is no statistical difference among STIP and spatio-temporal SIFT with both vocabulary sizes. In other words, SIFT can provide virtually the same recognition rates as STIP with a third of the vocabulary size. At first sight, this result suggests some advantage for the spatio-temporal SIFT, but such an assumption requires further examination.

A more detailed comparison between SIFT and STIP is pursued by computing some metrics related to computational complexity, which are shown in Table 6.2. As it can be seen, the STIP algorithm is much faster than SIFT and generates much less interest points also. This yields a considerably lower histogram computation time also for STIP[2].

[2]There are differences in PCA reduction and clustering times also, but since they are at least two orders of magnitude lower, they were disregarded in this analysis. The same for the SVM training and classification times.

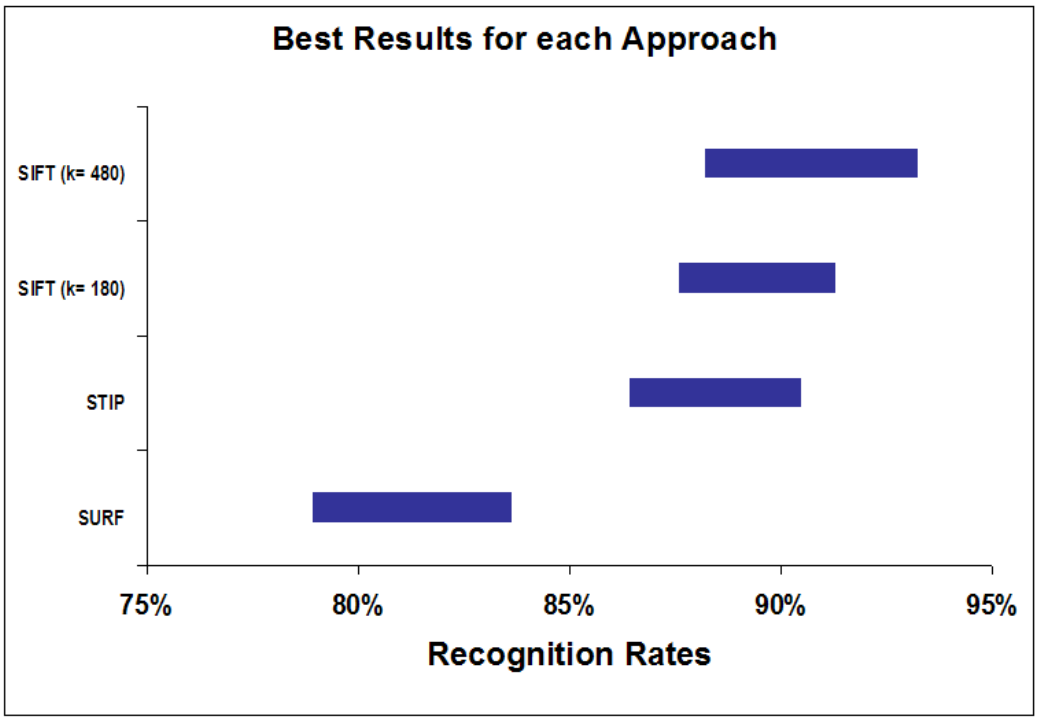


Figure 6.10. Comparing the best achieved results for each descriptor type, including both similar peaks found for SIFT.

In other words, although it is indeed possible to achieve a state-of-the art recognition rate on this database with a BoVF based on the SIFT descriptors applied to spatio-temporal frames, this result is achieved at a much higher complexity in time than with STIP. In addition, the apparent advantage of a smaller final BoVF representation in favor of spatio-temporal SIFT is somewhat unstable – can be seen in Figure 6.9 – and also very specific to this experimental setting.

Table 6.2. Comparison between spatio-temporal SIFT and STIP.

	ST-SIFT	**STIP**
Selection+Description Time	1327s	582s
Number of Points	504766	10886
Histogram Computation	1395s	113s

The complete set of comparisons (SIFT x SURF x STIP), together with the complimentary complexity analysis are published in Lopes et al. [2009c].

6.2.2 Concluding Remarks on Capturing Dynamics into BoVFs

The experimental results described in this section indicated that in terms of collecting dynamic information, STIP points present superior computational complexity than SIFT and SURF applied to spatio-temporal frames, although SIFT is able to achieve similar recognition rates.

It is worth pointing out, though, that while STIP is very focused in detecting sudden motion changes, SIFT is primarily designed to capture appearance. The Weizmann database presents few variations in appearance, and yet, SIFT was able to capture dynamics at the same level as STIP, when applied to spatio-temporal frames. So, it seems quite reasonable to investigate how they compare in a more realistic setting, in which appearance variations are also important.

6.3 Additional Applications

The development of a generic implementation for computing BoVF descriptors allowed for applications in two different scenarios, namely, nude detection – both in still images and videos – and detection of buildings in historical photographs. The work derived from those applications is summarized in the two next subsections.

6.3.1 Nude Detection

The ability to filter improper images by visual content instead of text has been the focus of several papers in the last years. Most of them start by detecting skin regions, and then applying some kind of shape or geometric analysis to recognize possible body postures which would be indicative of nudity or pornography.

In the pioneer work of Fleck et al. [1996], color and texture properties are combined to obtain a mask for skin regions, which are then grouped to match a human figure using geometric constraints derived from the human body structure. More recent examples of approaches combining skin detectors with geometrical or shape features are Zeng et al. [2004], Xu et al. [2005], Lee et al. [2006], Zhu et al. [2007b] and Liu et al. [2008a].

Some attempts at avoiding the need for a fine-tuned skin detection originated approaches based on generic color features, like in Yoo [2004], Belem et al. [2005] and Shih et al. [2007], for example. Visual features can also be combined with other feature types, as in the framework for recognizing pornographic web pages presented in Hu et al. [2007], in which text and image are both analyzed. In another multi-modal

approach, Zuo et al. [2008] propose a framework for recognizing pornographic movies by fusing the audio and video information.

The main problems with those proposals are the complexity of accurate skin detectors and the great variability in shapes and geometry of such images. Moreover, approaches based on skin detection are bound to fail when applied to monochromatic images.

In this work, we propose a BoVF-based approach using a variation of SIFT descriptor which includes color information called Hue-SIFT, described in van de Sande et al. [2008]. To the best of our knowledge, to the date we submitted the related papers, Deselaers et al. [2008] is the only work proposing a method also based on BoVFs to classify images into different categories of pornographic content. In their case, however, the features used to learn the vocabulary are the color values of patches around DoG interest points.

Nude Detection in Still Images

To compose a database for the evaluation of the proposed approach, a set of 180 images was collected from the web, and then manually classified. Examples of selected images are shown in Figure 6.11. In our experiments, SIFT and Hue-SIFT descriptors are extracted from all images, and then the vocabulary size and SVM model are extensively searched for, using a 5-fold cross validation scheme. For the vocabulary size (k), the experiments spawned values between 50 and 700. For each k value, the penalty of the SVM error term (C) is varied in a logarithmic scale from 10^{-5} to 10^{5}[3]. This procedure is repeated for SIFT and Hue-SIFT separately.

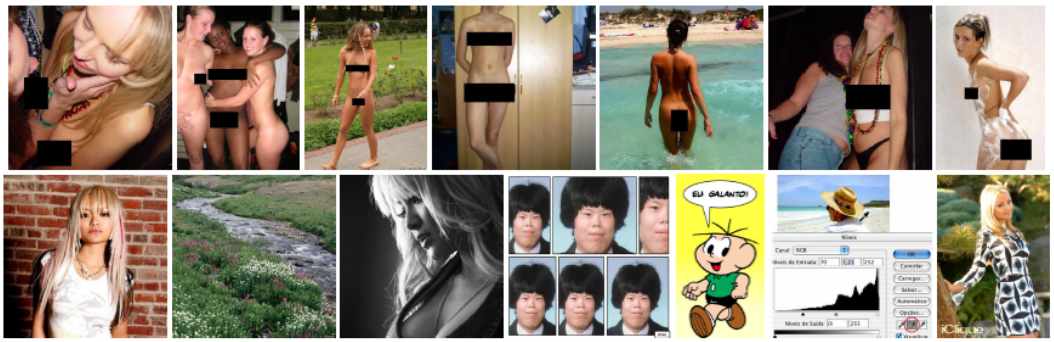


Figure 6.11. Some examples of nude and non-nude images collected for our evaluation database.

[3]The k and C values which achieved the best recognition rate were submitted to a finer search for C, but this procedure provided no significant enhancements

In order to evaluate the statistical significance of the differences found between the best recognition rates for SIFT and Hue-SIFT, ten new runs of the 5-fold cross validation are performed with fixed k and C. Folds composition changed randomly in each run, keeping the balance between nude and non-nude training examples. The average rates of each cross-validation run are used to compute confidence intervals for both results.

In Figure 6.12, the best recognition rates for all tested values of k are plotted both for Hue-SIFT and SIFT (the best recognition rate is the higher value achieved while varying penalty error term of the SVM model – the C parameter). It can be seen that SIFT recognition rates are consistently smaller for all tested vocabulary sizes, indicating the paramount importance of color information for nudity detection, which is in accordance with the basic implicit assumption of most skin-based proposals. The

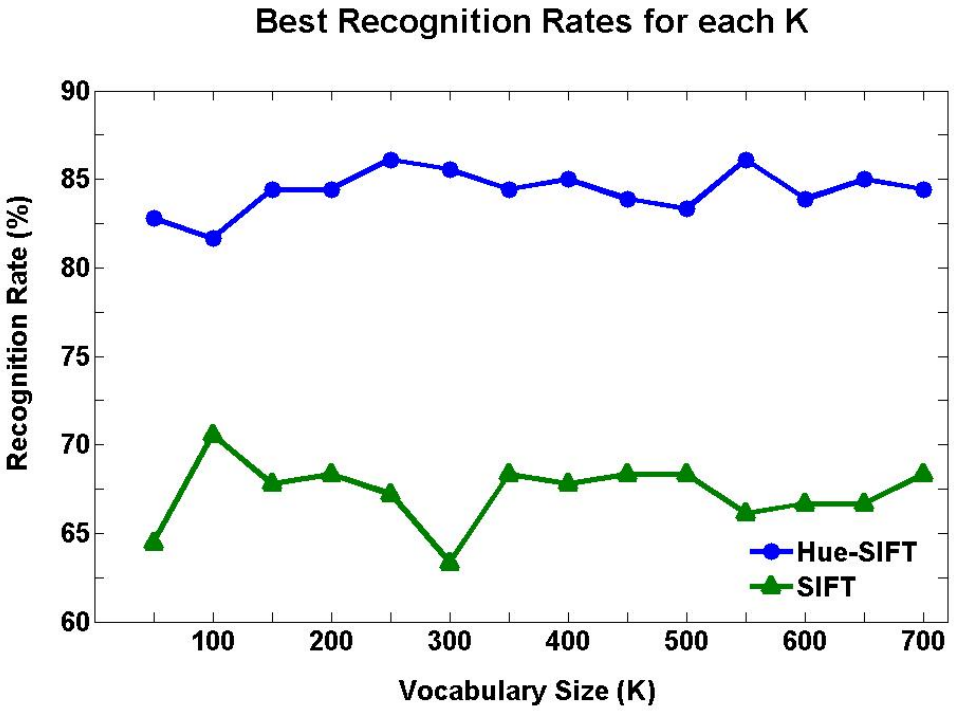


Figure 6.12. Comparing results based on SIFT and Hue-SIFT.

models which provided the peaks of the graph in Figure 6.12 – two for Hue-SIFT (at $k = 250$ and $k = 550$) and one for SIFT (at $k = 100$) – are used in the final experiment, which consists of ten runs of five-fold cross validations, randomly varying the folds composition. In Table 6.3, it is possible to see the final averages and their confidence intervals.

Unfortunately, the lack of standard annotated databases for nude or pornographic

Best Recognition Rates			
Descriptor	**Average (%)**	**Min. (%)**	**Max. (%)**
SIFT	64.8	61.7	68.0
Hue-SIFT	84.6	82.8	86.5

Table 6.3. Best recognition rates with Hue-SIFT and SIFT descriptors (confidence intervals – given by "Min." and "Max." values – have a confidence level of 95%).

images makes it difficult to compare results obtained by different approaches. Nevertheless, it is worth to mention that in a coarse comparison, our approach achieved an overall recognition rate very similar to those ones reported in recent literature. This was achieved without the usage of any skin and/or shape models and with the simplest possible SVM classifier. This result suggests that BoVF-based approaches are indeed a promising path to follow in the pursuit of effective filters for improper visual content.

The results described in this section were reported in Lopes et al. [2009a].

Nude Detection in Videos

The HueSIFT-based BoVF detailed above is used together with a voting scheme to identify nudity in video segments. The proposed voting scheme takes advantage of the existence of several similar frames in the video segments, as follows: firstly, the video segments are separated into frames, and some of them are selected according to a previously chosen sampling rate. The BoVFs for the selected frames are computed and then each frame is classified between nude and non-nude. For each video segment, every frame classification is counted as *votes* for that class, and the class receiving the higher number of votes is considered the video class. Because false-negatives are potentially more harmful in typical scenarios than false-positives, ties are considered as nude. Finally, in order to verify the recognition rates, the final video classification was compared to a manually built ground-truth.

Despite the importance of the nude detection task, there is no public available standard database for a significant comparison among different algorithms. Then, to be able to test the proposed algorithm, a collection of 179 video segments was created. Such segments were collected from 10 different movies and the entire segments database are available online[4].

[4] `http://www.npdi.dcc.ufmg.br/nudeDetection`

The experiments are designed to evaluate the ability of discriminating between nude and non-nude videos from BoVFs representations for the selected frames. Two different sample ratios were tested: 1/15 frame (2 frames per second) and 1/30 frames (1 frame per second). These sample ratios provided 2021 and 1011 frames to be classified, respectively. Sampling ratios lower than those ones were considered not desirable, to avoid missing short nude sequences between non-nude material.

Three different vocabulary sizes were tested for classification of those BoVF vectors between nude and non-nude: 60, 120 and 180. A linear SVM classifier is used, and its penalty error parameter (C) is refined by the same procedure described in Subsection 6.3.1. In order to compute the confidence intervals for the recognition rates of each k (and its best C previously determined), 30 new five-fold cross validation runs were performed, randomly varying the videos selected for each fold.

The recognition rates achieved in the above experiments are summarized in Table 6.4. In them, the recognition rate for frames (second column) are those expected if only a keyframe was selected to represent the entire video segment. In the third column, are the recognition rates using the proposed voting scheme. From those tables, it is possible to see that, in all cases, applying the voting algorithm causes a statistically significant increase in the overall recognition rate for the videos, when compared to the recognition rate of the individual frames. Such results indicate that, indeed, the voting scheme is able to take advantage of the existence of several similar frames to solve some dubious cases.

6.4 Concluding Remarks

This chapter describes the additional contributions derived of the exploratory work performed along the doctorate course, which provided the theoretical and practical basis to the development of this thesis. They are listed approximately in order of relevance, and derived publications are listed in Section 1.5.

Table 6.4. Comparing recognition rates for keyframe and voting based classification.

Voc. Size	Keyframe (%)	Voting (%)	Increase
60	76.4 ± 0.2	77.1 ± 0.4	0.7
120	80.2 ± 0.3	80.9 ± 0.4	0.7
180	83.9 ± 0.2	88.4 ± 0.6	4.5

(a) 1/30 frames

Voc. Size	Keyframe (%)	Voting (%)	Increase
60	79.1 ± 0.1	80.5 ± 0.4	1.4
120	83.7 ± 0.2	87.3 ± 0.4	3.6
180	85.9 ± 0.1	93.2 ± 0.4	7.3

(b) 1/15 frames

Chapter 7

Conclusion

In this thesis, we address the problem of action recognition in realistic videos based on their visual content only. More specifically, in our proposal we address two main issues:

- The inclusion of contextual information in the video descriptor to enhance classification.
- Dealing with the general lack of samples for training the action classifier.

In order to obtain a clear overview of the current state-of-the-art in the area, a broad survey on the subject was performed. This survey served two main purposes: firstly, it provided a firm foundation for a principled choice of the path to follow. Additionally, the survey showed a lack of previous surveys encompassing and properly organizing the most recent achievements in the area.

After having decided by a work based on BoVF representations and SVM classifiers, we firstly explored the issue of how to deal with the dynamic information in a BoVF-based approach for videos (Lopes et al. [2009d] and Lopes et al. [2009c]). Additionally, some scenarios other than action recognition were addressed (Lopes et al. [2009a], Lopes et al. [2009b] and Batista et al. [2009b]), providing further indications of the properness of BoVF-based approaches for challenging, realistic image and video databases. We also explored the cost-benefit both of the k-means algorithm and a less expensive, enhanced random vocabulary selection algorithm (Santos et al. [2010]).

Based on those preliminary results, we then proposed, implemented and evaluated a BoVF-based solution for action recognition in realistic videos from state-of-the art datasets. The main contribution of that solution can be summarized by the *addition of high-level contextual information by means of transfer learning*. The initial proposal

was based on some indications coming from different recent works in the area, which supported the underpinning assumptions of the proposed methodology, which are listed and justified in Chapter 4.

We advocate that the current approach for action recognition presents a number of advantages over existing ones, namely:

1. it relies on BoVF-based concept recognition instead of accurate detectors or global low-level features, therefore being more robust to several kinds of visual variations.

2. it gathers contextual information from cleanly, instead of noisy, annotated data.

3. the solution, based on TL theory, is generic enough to encompass any number of external databases, drawing on and adding value to previous annotation efforts of the academic community.

4. the context description is kept in the semantic level, contributing to narrow the semantic gap for the specific case of human action recognition.

7.1 Future Work

The next natural step to this work is to apply the solution proposed in a number of databases of realistic action videos in fully-comparable experiment settings. In particular, we propose that a number of issues deserve some attention in these near future investigations:

- How the number of concepts affects the enhancement in action recognition by transfer? More is always better or is there a *plateau* above which no enhancement is possible?

- Which actions benefit more from transfer and why? Could we devise ways of anticipating that behavior?

- How to explore the relation between concepts and actions more explicitly?

- How different auxiliary databases could be used in synergy?

In a more long-term vein, it is worth noticing that spite of the success of BoVF-based approaches, a number of limitations are yet to be overcome by them in order to achieve solutions that are mature enough to be incorporated in real-world tools. One of those issues is the *ad-hoc* nature of the visual vocabulary learning process. Although

some papers discussed (Section 3.4.2) indicate better alternatives to pure k-means, there is no principled methodology neither to build an optimal vocabulary, nor even to preemptively assess the vocabulary quality, given a specific database.

The relatively small number of proposals dealing with local descriptors in alternative ways (Section 3.4.2) prevents the anticipation of specific trends in this direction, but it is noticeable that both BoVF and non-BoVF-based approaches have been moved from an initial preference for sparse set of interest points to a current trend towards denser sets, on the assumption that they are more appropriate for realistic scenarios. This premise is reinforced, for example, by the comparison among several 3D point detectors and descriptors performed by Wang et al. [2009], which concluded that, except for the KTH database – which has very little contextual information, given its neutral background – regularly spaced dense samplings of points perform better at action recognition than interest points.

Approaches like ours, based on concept probabilities are somewhat underexplored for action recognition, mainly if one considers their widespread usage in CBVR systems. One reason for this apparent lack of interest might be the scarcity of annotated training samples for each concept classifier, as well as the issues raised by the usage of meta-classification and classification fusion. We addressed them with the aid of the theory of Transfer Learning (TL), but we consider there is abundant space for alternative approaches.

The scarcity of labeled data for action recognition itself is an important issue which needs to be tackled in order to allow more significant advances in the area. Some initiatives have generated *de facto* standard databases, like Weizmann Blank et al. [2005] and KTH Schuldt et al. [2004] controlled databases, and, recently, more realistic ones, like the Hollywood Movies Datasets (Laptev et al. [2008] – Hollywood, Marszalek et al. [2009] – Hollywood2) and the Action Dataset Liu et al. [2009]. Such annotation efforts have been fundamental to research advances achieved in the area in the last few years, but they suffer the limitation of being somewhat isolated efforts and therefore, necessarily limited in size and scope. The TRECVID benchmark, which is well-known for its collective annotation efforts at larger scales, has had an event detection/recognition track for a few years, but videos and ground-truth data are available for participants only.

Although most current approaches for action recognition focus on enhancing recognition capability, it is worth mentioning that in order to scale up to the dimensions of the web or of any large database, efficiency needs to be taken into account. Some efforts in this direction are available in the literature. In the work of Lin et al. [2009], for example, lookup tables of action prototypes are used to speed-up action

classification. In Uemura et al. [2008], Ji et al. [2008] and Mikolajczyk and Uemura [2008], the features are quantized using vocabulary trees, which lead to more efficient matching when compared with traditional codebooks. In the realm of more typical BoVF schemes, compact vocabularies – like those in Liu and Shah [2008], Jiang and Ngo [2009] and Liu et al. [2009] – can help to reduce the overall computational effort. In Ke et al. [2005], Ke et al. [2007b] and Laptev and Perez [2007], efficient boosting algorithms are applied for volumetric matching. In Mikolajczyk and Uemura [2008], the high computational complexity of the proposed approach – based on several dense interest point detectors – is explicitly pointed out, and leaded the authors to implement their recognition framework in a parallel architecture. Yet, action recognition by visual information still require a huge computational effort, unable to compete with pure textual solutions.

In other words, this research area is still plenty of challenges both in terms of recognition ability and performance in order to be merged with current generic search engines. In spite of that, some specific applications can already be developed using existing state-of-the-art approaches.

Finally, we firmly believe that a full mature solution to indexing and search of multimedia content is going to be multi-modal, including not only textual, visual and motion information, but also audio, personal and social evidences, all fused together.

Bibliography

Abdelkader, M. F., Roy-Chowdhury, A. K., Chellappa, R., and Akdemir, U. (2008). Activity representation using 3d shape models. *J. Image Video Process.*, 8(2):1–16.

Agarwal, S. and Awan, A. (2004). Learning to detect objects in images via a sparse, part-based representation. *TPAMI*, 26(11):1475–1490. Member-Dan Roth.

Aggarwal, J. and Cai, Q. (1997). Human motion analysis: a review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102.

Aggarwal, J. K. and Park, S. (2004). Human motion: Modeling and recognition of actions and interactions. In *3DPVT '04*, pages 640–647, Washington, DC, USA. IEEE Computer Society.

Ahad, M. A. R., Tan, J., Kim, H., and Ishikawa, S. (2008). Human activity recognition: Various paradigms. In *ICCAS '08*, pages 1896–1901.

Ahammad, P., Yeo, C., Ramchandran, K., and Sastry, S. (2007). Unsupervised discovery of action hierarchies in large collections of activity videos. In *MMSP '07*, pages 410–413.

Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. Technical report, Rochester, NY, USA.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Batista, N. C., Lopes, A. P. B., and de Albuquerque Araújo, A. (2009a). Detecção de edifícios em fotografias históricas utilizando vocabulários visuais. In *Proceedings of CLEI '09*.

Batista, N. C., Lopes, A. P. B., and de Albuquerque Araújo, A. (2009b). Detecting buildings in historical photographs using bag-of-keypoints (in press). In *Proceedings of SIBGRAPI '09*. IEEE Computer Society.

Belem, R. J. S., B., J. M., de Moura Cavalcanti, E. S., and Nascimento, M. A. (2005). SNIF: A simple nude image finder. In *Proceedings of the Third Latin American Web Congress (LA-WEB)*, pages 252--258, Washington, DC, USA. IEEE Computer Society.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *ICCV05*, volume II, pages 1395–1402.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *TPAMI*, 23(3):257–267.

Bobick, A. F., Intille, S. S., Davis, J. W., Baird, F., Pinhanez, C. S., Campbell, L. W., Ivanov, Y. A., Schütte, A., and Wilson, A. (1999). The kidsroom: A perceptually-based interactive and immersive story environment. *Presence: Teleoper. Virtual Environ.*, 8(4):369–393.

Branzan Albu, A., Beugeling, T., Virji Babul, N., and Beach, C. (2007). Analysis of irregularities in human actions with volumetric motion history images. In *Motion '07*, page 16.

Bregonzio, M., Gong, S., and Xiang, T. (2009). Recognising action as clouds of space-time interest points. In *CVPR '09*, pages 1948–1955.

Buehler, P., Zisserman, A., and Everingham, M. (2009). Learning sign language by watching tv (using weakly aligned subtitles). In *CVPR '09*, pages 2961–2968.

Buxton, H. (2003). Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, 21(1):125 – 136.

Campbell, C. (2008). Introduction to support vector machines. Video Lectures – http://videolectures.net.

Cedras, C. and Shah, M. (1995). Motion-based recognition: A survey. *IVC*, 13(2):129–155.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.* Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapelle, O., Haffner, P., and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *TNN*, 10(5):1055–1064.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR '09*, pages 1932–1939.

Chellappa, R., Roy-Chowdhury, A. K., and Zhou, S. K. (2005). Recognition of humans and their activities using video. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 1(1):1–173.

Cooper, H. and Bowden, R. (2007). Sign language recognition using boosted volumetric features. In *Proc. IAPR Conf. on Machine Vision Applications*, pages 359–362.

Cristianini, J. S.-T. . N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Cuntoor, N. (2006). Morse functions for activity classification using spatiotemporal volumes. In *BP06*, page 20.

Cuntoor, N., Yegnanarayana, B., and Chellappa, R. (2008). Activity modeling using event probability sequences. *IP*, 17(4):594–607.

Dai, W., Yang, Q., Xue, G., and Yu, Y. (2007). Boosting for transfer learning. In Ghahramani, Z., editor, *ICML '07*, pages 193--200. Omnipress.

Datong, C., Wactlar, H., yu Chen, M., Can, G., Bharucha, A., and Hauptmann, A. (2008). Recognition of aggressive human behavior using binary local motion descriptors. In *EMBS '08*, pages 5238–5241.

Davis, J. and Domingos, P. (2009). Deep transfer via second-order Markov logic. In Bottou, L. and Littman, M., editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 217--224, Montreal. Omnipress.

de Avila, S. E. F., Lopes, A. P. B. a., da Luz, Jr., A., and de Albuquerque Araújo, A. (2011). Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recogn. Lett.*, 32:56--68.

Deselaers, T., Pimenidis, L., and Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. In *International Conference on Pattern Recognition (ICPR)*, Florida, USA.

Dhillon, P., Nowozin, S., and Lampert, C. (2009). Combining appearance and motion for human action classification in videos. In *Proceedings of IEEE CVPRW '09*, volume 0, pages 22–29.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *ICCCN '05*, pages 65–72, Washington, DC, USA. IEEE Computer Society.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., and Ponce, J. (2009). Automatic annotation of human actions in video. In *ICCV '09*.

Ebadollahi, S., Xie, L., fu Chang, S., and Smith, J. (2006a). Visual event detection using multi-dimensional concept dynamics. pages 881–884, Los Alamitos, CA, USA. IEEE Computer Society.

Ebadollahi, S., Xie, L., fu Chang, S., and Smith, J. (2006b). Visual event detection using multi-dimensional concept dynamics. In *ICME '06*, volume 0, pages 881–884, Los Alamitos, CA, USA. IEEE Computer Society.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *ICCV03*, page 726, Washington, DC, USA. IEEE Computer Society.

Ermis, E., Saligrama, V., Jodoin, P., and Konrad, J. (2008). Motion segmentation and abnormal behavior detection via behavior clustering. In *ICIP '08*, pages 769–772.

Fathi, A. and Mori, G. (2008). Action recognition by learning mid-level motion features. In *CVPR08*, pages 1–8.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524--531.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Filipovych, R. and Ribeiro, E. (2008). Learning human motion models from unsegmented videos. In *CVPR '08*, pages 1–7.

Fleck, M. M., Forsyth, D. A., and Bregler, C. (1996). Finding naked people. In *Proceedings of the 4th European Conference on Computer Vision-Volume II (ECCV)*, pages 593--602, London, UK. Springer-Verlag.

Gavrila, D. M. (1999). The visual analysis of human movement: a survey. *Comput. Vis. Image Underst.*, 73(1):82--98.

Gilbert, A., Illingworth, J., and Bowden, R. (2008). Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV '08*, pages 222--233, Berlin, Heidelberg. Springer-Verlag.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *TPAMI*, 29(12):2247–2253.

Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 object category dataset. Technical report 7694, California Institute of Technology.

Hatun, K. and Duygulu, P. (2008). Pose sentences: A new representation for action recognition using sequence of pose words. In *ICPR '08*, pages 1–4.

Haubold, A. and Naphade, M. (2007). Classification of video events using 4-dimensional time-compressed motion features. In *CIVR '07*, pages 178--185, New York, NY, USA. ACM.

Herbert Baya, Andreas Ess, T. T. and Gool, L. V. (2008). Speed-up robust features (surf). *Computer Vision and Image Understanding*, 110:346–359.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425.

Hu, M., Ali, S., and Shah, M. (2008a). Detecting global motion patterns in complex videos. In *ICPR '08*, pages 1–5.

Hu, M., Ali, S., and Shah, M. (2008b). Learning motion patterns in crowded scenes using motion flow field. In *ICPR '08*, pages 1–5.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *SMC-C*, 34(3):334–352.

Hu, W., Wu, O., Chen, Z., and Fu, Z. (2007). Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1019--1034.

Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., and Huang, T. (2009). Action detection in complex scenes with spatial and temporal ambiguities. In *Proceedings of IEEE ICCV '09*.

Ji, R., Sun, X., Yao, H., Xu, P., Liu, T., and Liu, X. (2008). Attention-driven action retrieval with dtw-based 3d descriptor matching. In *MM '08*, pages 619--622, New York, NY, USA. ACM.

Jiang, H., Drew, M., and Li, Z. (2006). Successive convex matching for action detection. In *CVPR06*, volume II, pages 1646–1653.

Jiang, Y.-G. and Ngo, C.-W. (2009). Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Comput. Vis. Image Underst.*, 113(3):405--414.

Jiang, Y.-G., Ngo, C.-W., and Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of ACM CIVR '07*, pages 494–501.

Junejo, I., Dexter, E., Laptev, I., and Perez, P. (2008). Cross-view action recognition from temporal self-similarities. In *ECCV '08*, pages II: 293–306.

Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 604–610.

Kaiser, B. and Heidemann, G. (2008). Qualitative analysis of spatio-temporal event detectors. In *ICPR '08*, pages 1–4.

Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N., Roy Chowdhury, A., Kruger, V., and Chellappa, R. (2004). Identification of humans using gait. *IP*, 13(9):1163–1173.

Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *ICCV '05*, pages 166--173, Washington, DC, USA. IEEE Computer Society.

Ke, Y., Sukthankar, R., and Hebert, M. (2007a). Event detection in crowded videos. In *ICCV '07*, pages 1–8.

Ke, Y., Sukthankar, R., and Hebert, M. (2007b). Spatio-temporal shape and flow correlation for action recognition. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Kennedy, L. (2006). Revision of LSCOM Event/Activity Annotations, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. Technical report, Columbia University.

Kennedy, L. and Hauptmann, A. (2006). LSCOM Lexicon Definitions and Annotations(Version 1.0). Technical report, Columbia University.

Klaser, A., Marszalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC '08*.

Kosta, G., Pedro, C., and Benoit, M. (2006). Modelization of limb coordination for human action analysis. In *Image Processing, 2006 IEEE International Conference on*, pages 1765–1768.

Kruger, V., Kragic, D., Ude, A., and Geib, C. (September 2007). The meaning of action: a review on action recognition and mapping. *Advanced Robotics*, 21:1473–1501(29).

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *ICCV '09*.

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107--123.

Laptev, I., Caputo, B., Schuldt, C., and Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *CVIU*, 108(3):207–229.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *ICCV03*, pages 432–439.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *CVPR '08*, pages 1–8.

Laptev, I. and Perez, P. (2007). Retrieving actions in movies. In *ICCV '07*, pages 1–8.

Lavee, G., Khan, L., and Thuraisingham, B. (2007). A framework for a video analysis tool for suspicious event detection. *Multimedia Tools Appl.*, 35(1):109–123.

Lay, D. C. (2002). *Linear algebra and its applications*. Addison-Wesley, New York, 3rd edition.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR06*, volume II, pages 2169–2178.

Lee, J.-S., Kuo, Y.-M., and Chung, P.-C. (2006). The adult image identification based on online sampling. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 2566--2571. IEEE.

Li, X. (2007). Hmm based action recognition using oriented histograms of optical flow field. *Electronics Letters*, 43(10):560–561.

Lin, Z., Jiang, Z., and Davis, L. S. (2009). Recognizing actions by shape-motion prototype trees. In *Proceedings of ICCV '09*.

Liu, B.-B., Su, J.-Y., Lu, Z.-M., and Li, Z. (2008a). Pornographic images detection based on CBIR and skin analysis. *International Conference on Semantics, Knowledge and Grid (SKG)*, 0:487–488.

Liu, J., Ali, S., and Shah, M. (2008b). Recognizing human actions using multiple features. *CVPR '08*, pages 1–8.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos 'in the wild'. In *Proceedings of IEEE CVPR '09*, pages 1996–2003.

Liu, J. and Shah, M. (2008). Learning human actions via information maximization. In *"CVPR '08"*.

Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., and de Albuquerque Araújo, A. (2009a). A bag-of-features approach based on hue-sift descriptor for nude detection. In *Proceedings of EUSIPCO '09*.

Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., de Miranda Coelho, M., and de Albuquerque Araújo, A. (2009b). Nude detection in video using bag-of-visual-features. In *Proceedings of SIBGRAPI '09*. IEEE Computer Society.

Lopes, A. P. B., Oliveira, R. S., de Almeida, J. M., and de Albuquerque Araújo, A. (2009c). Comparing alternatives for capturing dynamic information in bag of visual features approaches applied to human actions recognition. In *Proceedings of MMSP '09*.

Lopes, A. P. B., Oliveira, R. S., de Almeida, J. M., and de Albuquerque Araújo, A. (2009d). Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. In *Proceeding of SIBGRAPI '09*, pages 1–7. IEEE Computer Society.

Lopes, A. P. B., Santos, E. R. d. S., do Valle Jr., E. d. A., de Almeida, J. M., and Araújo, A. d. A. (2011). Transfer learning for human action recognition. In *Proceedings of SIBGRAPI '11*.

Lowe, D. (1999). Object recognition from local scale-invariant features. *ICCV '99*, 2:1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91--110.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of IJCAI '81*, pages 674--679.

Ma, Y. and Cisar, P. (2009). Event detection using local binary pattern based dynamic textures. volume 0, pages 38–44.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *CVPR '09*, pages 2929–2936.

Messing, R., Pal, C., and Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *Proceedings of ICCV '09*.

Mikolajczyk, K. and Uemura, H. (2008). Action recognition with motion-appearance vocabulary forest. In *CVPR '08*, pages 1–8.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126.

Mokhber, A., Achard, C., and Milgram, M. (2008). Recognition of human behavior by space-time silhouette characterization. *Pattern Recogn. Lett.*, 29(1):81–89.

Moreno, P. J., Ho, P. P., and Vasconcelos, N. (2003). A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Proceedings of NIPS '03*, Vancouver, Canada.

Naphade, M. R., Kennedy, L., Kender, J., Chang, S. F., Smith, J. R., Over, P., and Hauptmann, A. (2005). A light-scale concept ontology for multimedia understanding for trecvid 2005. Technical report, IBM Research.

Niebles, J. and Li, F. (2007). A hierarchical model of shape and appearance for human action classification. In *CVPR '07*, pages 1–8.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318.

Ning, H., Han, T. X., Walther, D. B., Liu, M., and Huang, T. S. (2009). Hierarchical space-time model enabling efficient search for human actions. *Trans. Cir. and Sys. for Video Technol.*, 19(6):808--820.

Ning, H., Hu, Y., and Huang, T. (2007). Searching human behaviors using spatial-temporal words. In *Proceedings of the IEEE International Conference on Image Processing*, pages 337–340.

Nowozin, S., Bakir, G., and Tsuda, K. (2007). Discriminative subsequence mining for action classification. *ICCV '07*, pages 1–8.

Oikonomopoulos, A., Patras, I., and Pantic, M. (2005). Spatiotemporal saliency for human action recognition. In *ICME05*, pages 1–4.

Oikonomopoulos, A., Patras, I., and Pantic, M. (2006). Kernel-based recognition of human actions using spatiotemporal salient points. In *V4HCI06*, page 151.

Oikonomopoulos, A., Patras, I., and Pantic, M. (2009). An implicit spatiotemporal shape model for human activity localization and recognition. In *Proceedings of IEEE CVPR-4HB '09*, pages 27–33.

Over, P., Ianeva, T., Kraaij, W., and Smeaton, A. F. (2005). Trecvid 2005 - an overview. In *In Proceedings of TRECVID 2005*.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *Transactions on Knowledge and Data Engineering (pre-print)*.

Papadopoulos, G., Mezaris, V., Kompatsiaris, I., and Strintzis, M. (2008). Estimation and representation of accumulated motion characteristics for semantic event detection. In *ICIP '08*, pages 41–44.

Pavlovic, V. I., Sharma, R., and Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE PAMI*, 19:677--695.

Peursum, P., Bui, H. H., Venkatesh, S., and West, G. (2004). Human action segmentation via controlled use of missing data in hmms. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, pages 440–445, Washington, DC, USA. IEEE Computer Society.

Peursum, P., Bui, H. H., Venkatesh, S., and West, G. (2005). Robust recognition and segmentation of human actions using hmms with missing observations. *EURASIP J. Appl. Signal Process.*, 2005(1):2110–2126.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.*, 108(1-2):4–18.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Comput.*, 28(6):976--990.

Rahman, M. M. and Ishikawa, S. (2005). Human motion recognition using an eigenspace. *Pattern Recogn. Lett.*, 26(6):687–697.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. (2007). Self-taught learning: transfer learning from unlabeled data. pages 759--766.

Rapantzikos, K., Avrithis, Y., and Kollias, S. (2009). Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR '09*, pages 1454–1461.

Ren, H. and Xu, G. (2002). Human action recognition in smart classroom. In *AFGR02*, pages 399–404.

Ren, W., Singh, S., Singh, M., and Zhu, Y. S. (2009). State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recogn.*, 42(2):267–282.

Ryoo, M. S. and Aggarwal, J. K. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Proceedings of ICCV '09*.

Santos, E. R. d. S., Lopes, A. P. B., do Valle Jr., E. d. A., de Almeida, J. M., and Araújo, A. d. A. (2010). Vocabulários visuais para recuperação de informação multimídia. In *Proceedings of WEBMEDIA '10*.

Savarese, S., DelPozo, A., Niebles, J., and Fei-Fei, L. (2008). Spatial-temporal correlatons for unsupervised action classification. In *WMVC '08*, pages 1–8.

Savarese, S., Winn, J., and Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *CVPR '06*, pages 2033–2040, Washington, DC, USA. IEEE Computer Society.

Schindler, G., Zitnick, L., and Brown, M. (2008). Internet video category recognition. In *InterNet '08*, pages 1–7.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *ICPR '04*, volume III, pages 32–36.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *MULTIMEDIA '07*, pages 357–360, New York, NY, USA. ACM.

Seo, H. J. and Milanfar, P. (2009). Detection of human actions from a single example. In *CVPR '09*, pages 1965 –1970.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *Trans. Pattern Anal. Mach. Intell.*, 29(3):411--426.

Shah, M. (2003). Understanding human behavior from motion imagery. *Mach. Vision Appl.*, 14(4):210–214.

Shechtman, E. and Irani, M. (2005). Space-time behavior based correlation. In *CVPR '05*, pages 405–412, Washington, DC, USA. IEEE Computer Society.

Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *Proceedings of IEEE CVPR '07*, pages 1 –8.

Shih, J.-L., Lee, C.-H., and Yang, C.-S. (2007). An adult image identification system employing image retrieval technique. *Pattern Recognition Letters*, 28(16):2367--2374.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349--1380.

Snoek, C. G. M. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl.*, 25(1):5--35.

Snoek, C. G. M. and Worring, M. (2008). Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322.

Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., and Li, J. (2009a). Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2004–2011.

Sun, X., Chen, M., and Hauptmann, A. (2009b). Action recognition via local descriptors and holistic features. In *Proceedings of IEEE CVPR-4HB '09*, pages 58–65.

Tan, R. and Davis, J. W. (2004). Differential video coding of face and gesture events in presentation videos. *Comput. Vis. Image Underst.*, 96(2):200–215.

Thurau, C. and Hlavac, V. (2008). Pose primitive based human action recognition in videos or still images. In *CVPR '08*, pages 1–8.

Tien, M.-C., Wang, Y.-T., Chou, C.-W., Hsieh, K.-Y., Chu, W.-T., and Wu, J.-L. (2008). Event detection in tennis matches based on video data mining. In *Proceedings of IEEE ICME '08*, pages 1477 –1480.

Tong, X., Duan, L., Xu, C., Tian, Q., and Lu, H. (2006). Local motion analysis and its application in video based swimming style recognition. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 1258--1261, Washington, DC, USA. IEEE Computer Society.

Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O. (2008). Machine recognitionnology of human activities: A survey. *Trans. Circuits and Sys. for Video Tech.*, 18(11):1473–1488.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280.

Uemura, H., Ishikawa, S., and Mikolajczyk, K. (2008). Feature tracking and motion compensation for action recognition. In *Proceedings of BMVA BMVC '08*, pages 1–8.

Ulges, A., Schulze, C., Koch, M., and Breuel, T. M. (2009). Learning automatic concept detectors from online video. *Computer Vision and Image Understanding*, In Press, Corrected Proof:–.

Ullah, M. M., Parizi, S. N., and Laptev, I. (2010). Improving bag-of-features action recognition with non-local cues. In *BMVC '10*, pages 95.1--11. doi:10.5244/C.24.95.

van de Sande, K., Gevers, T., and Snoek, C. (2008). Evaluation of color descriptors for object and scene recognition. In *CVPR '08*, Los Alamitos, CA, USA. IEEE Computer Society.

van de Sande, K., Gevers, T., and Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1582--1596.

van de Weijer, J., Gevers, T., and Bagdanov, A. D. (2006). Boosting color saliency in image feature detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:150--156.

Viitaniemi, V. and Laaksonen, J. (2008). Experiments on selection of codebooks for local image feature histograms. In *Proceedings of the 10th international conference*

on Visual Information Systems: Web-Based Visual Information Search and Management, VISUAL '08, pages 126--137, Berlin, Heidelberg. Springer-Verlag.

Viola, P. and Jones, M. (2001). Robust real-time object detection. In *Int. Journal of Comp. Vision*.

Wang, F., Jiang, Y.-G., and Ngo, C.-W. (2008). Video event detection using motion relativity and visual relatedness. In *MM '08*, pages 239--248, New York, NY, USA. ACM.

Wang, H., Ullah, M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC '09*, pages 1–5.

Wang, J. and Singh, S. (2003). Video analysis of human dynamics: A survey. *Real-TimeImg*, 9(5):320–345.

Wang, L., Hu, W., and Tan, T. (2003). Recent developments in human motion analysis. *PR*, 36(3):585–601.

Wang, Y., Sabzmeydani, P., and Mori, G. (2007). Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *HUMO07*, pages 240–254.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV '08*, pages 650--663, Berlin, Heidelberg. Springer-Verlag.

Wong, S.-F. and Cipolla, R. (2007). Extracting spatiotemporal interest points using global information. In *ICCV '07*, pages 1–8.

Wong, S.-F., Kim, T.-K., and Cipolla, R. (2007). Learning motion categories using both semantic and structural information. In *CVPR '07*, pages 1--6.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975--1005.

Xie, L., Xu, P., Chang, S.-F., Divakaran, A., and Sun, H. (2004). Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recogn. Lett.*, 25(7):767--775.

Xu, D. and Chang, S.-F. (2008). Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1985--1997.

Xu, Y., Li, B., Xue, X., and Lu, H. (2005). Region-based pornographic image detection. *IEEE 7th Workshop on Multimedia Signal Processing (MMSP)*, pages 1--4.

Yang, J., Jiang, Y.-G., Hauptmann, A. G., and Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval (MIR)*, pages 197--206, New York, NY, USA. ACM.

Yanik, P. M., Manganelli, J., Merino, J., Smolentzov, L., Walker, I. D., Brooks, J. O., and Green, K. E. (2011). Sensor placement for activity recognition: comparing video data with motion sensor data. *International Journal of Circuits, Systems and Signal Processing*, 5:279--286.

Yilmaz, A. and Shah, M. (2005). Actions sketch: a novel action representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 984–989.

Yoo, S.-J. (2004). Intelligent multimedia information retrieval for identifying and rating adult images. In *Proceedings of 8th International Conference Knowledge-Based Intelligent Information and Engineering Systems (KES)*, volume 3213 of *Lecture Notes in Computer Science*, pages 164--170. Springer.

Zelnik-Manor, L. (2006). Statistical analysis of dynamic actions. *TPAMI*, 28(9):1530–1535. Member-Michal Irani.

Zelnik-Manor, L. and Irani, M. (2001). Event-based analysis of video. In *CVPR2001*, volume 2, pages 123–130.

Zeng, W., Gao, W., Zhang, T., and Liu, Y. (2004). Image guarder: An intelligent detector for adult images. In *Asian Conference on Computer Vision*, pages 198--203, Jeju Island, Korea.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007a). Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213--238.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007b). Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213--238.

Zhao, Z. and Elgammal, A. (2008). Human activity recognition from frame's spatiotemporal representation. In *ICPR08*, pages 1–4.

Zhou, X., Zhuang, X., Yan, S., Chang, S.-F., Hasegawa-Johnson, M., and Huang, T. S. (2008). Sift-bag kernel for video event analysis. In *MM '08*, pages 229–238, New York, NY, USA. ACM.

Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., and Yao, H. (2007a). Trajectory based event tactics analysis in broadcast sports video. In *MULTIMEDIA '07*, pages 58–67, New York, NY, USA. ACM.

Zhu, H., Zhou, S., Wang, J., and Yin, Z. (2007b). An algorithm of pornographic image detection. In *Proceedings of the Fourth International Conference on Image and Graphics (ICIG)*, pages 801--804, Washington, USA. IEEE Computer Society.

Zuo, H., Wu, O., Hu, W., and Xu, B. (2008). Recognition of blue movies by fusion of audio and video. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 37--40.

—